ARL-TR-7466 • SEP 2015

US Army Research Laboratory

Effects of Agent Transparency on Multi-Robot Management Effectiveness

by Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Daniel Barber, Katelyn Procci, and Michael Barnes

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

ARL-TR-7466 ● SEP 2015

US Army Research Laboratory

Effects of Agent Transparency on Multi-Robot Management Effectiveness

by Joseph E Mercado
Oak Ridge Associated Universities, Oak Ridge, TN

Jessie YC Chen and Michael Barnes
Human Research and Engineering Directorate, ARL

Michael A Rupp, Daniel Barber, and Katelyn Procci
University of Central Florida, Orlando, FL

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED (From - To)**
September 2015	Final	October 2013–September 2014

4. TITLE AND SUBTITLE Effects of Agent Transparency on Multi-Robot Management Effectiveness	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Daniel Barber, Katelyn Procci, and Michael Barnes	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) US Army Research Laboratory ATTN: RDRL-HRM-AR Aberdeen Proving Ground, MD 21005-5425	**8. PERFORMING ORGANIZATION REPORT NUMBER** ARL-TR-7466
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The objective of the study was to investigate the effects of agent transparency on operator performance in the context of joint human-agent decision making in multi-robot management. The agent display configurations were based on the 3 levels of the situation awareness–based agent transparency model (basic-information, reasoning, and projections/uncertainty). Results showed that participants calibrated their trust in the agent more effectively (proper reliance and correct rejections) and reported higher levels of trust when they were provided with the agent's reasoning and uncertainty information. No speed-accuracy trade-offs were observed. Nor did the participants report higher levels of workload when agent transparency increased. Working memory capacity was found to be a significant predictor of participants' trust in the agent. Individual differences in spatial ability accounted for variations in ocular indices of workload across display configurations.

15. SUBJECT TERMS

human-robot interaction, autonomous systems, transparency, trust, situation awareness, SA

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Jessie Chen
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	102	**19b. TELEPHONE NUMBER (Include area code)** 407-384-5435

Contents

List of Figures

List of Tables

Acknowledgments

We would like to thank Jonathan Harris, Isacc Yi, Gloria Calhoun, Mark Draper, Andre Garcia, and Erica Valiente for their contribution to this project.

1. Introduction

Agility in tactical decision making, mission management, and control is the key attribute for enabling heterogeneous multi-unmanned vehicle (UxV) teams to successfully manage the “fog of war” with its inherent complex, ambiguous, and time-challenged conditions. Mission effectiveness will rely on rapid identification and management of uncertainties that can disrupt an autonomous team’s ability to complete complex operations safely. As a result, many of today’s operators use complex human-machine systems on a daily basis. Further, as operators have to plan and direct multiple UxVs simultaneously to achieve mission objectives, human operators are often not able to maintain efficient and effective performance (Chen and Barnes 2012a). These decrements, which may lead to both mission failure and loss of life and property, may partially stem from the high information flow rate required to supervise multiple UxVs concurrently (Paas and Merriënboer 1994). Thus it is necessary to lower the cognitive load placed on the operator (Hwang et al. 2008) by presenting appropriate information when needed (Lyons and Havig 2014).

To decide what information is presented to the operator, intelligent agents (IAs) have been created to perform the role of an intermediary between the operator and each individual unmanned vehicle. In artificial intelligence, the concept of an agent is defined as anything that has the ability to perceive its environment through sensors and act upon its environment (Russel and Norvig 2009). An IA is an agent that has some level of autonomy, meaning it can act with limited authority from others and is responsible for reaching decisions (Russel and Norvig 2009). We specifically use the term IA to denote a software agent that is incorporated into a human machine system for the purpose of shared decision making between the IA and the system’s operator (e.g., Chen et al. 2014). Thus, instead of manually issuing commands to each UxV, the operator acts as a supervisor receiving feedback from and providing instructions to an IA who relays these commands to the UxVs to accomplish their shared mission (Chen and Barnes 2012b).

In this design the human operator always has the ultimate decision authority, which is an example of mixed-initiative decision making as defined by Goodrich (2010). However, with increasing levels of autonomy, human operators may not understand the information provided by the IA (i.e., generating automated plans) due to the operators’ difficulties understanding the rationale behind the decision-making processes (Linegang et al. 2006). This lack of understanding may lead to disuse or overreliance of the system (Parasuraman and Riley 1997). To alleviate this problem and facilitate fluid mixed-initiative decision making between the human operator and the IA, the agent-user interface must support optimal transparency, conveying the rationale behind its recommendations without burdening the operator with an overwhelming amount of data (Lee and See 2004). One approach that has been used in the literature with success has been the Playbook architecture (Miller et al. 2004). In this technique the human operator acts similarly to a coach of a sports team who conveys his or her goals and directs specific behaviors by calling a particular play (a specified command that conveys specific behaviors to be completed) and the UxVs act as the “players” who autonomously carry out the instructions contained in the play.

1.1 Automation Transparency

The 3 most common challenges for humans interacting with highly automated systems is understanding the current system state, comprehending the reasons for its current behavior, and projecting what its next behavior will be (Sarter and Woods 1995). In response to those 3 critical questions, transparency in automated systems has become a critical research question. Agent transparency is the IA’s ability to communicate information to the human operator in a clear and efficient manner, which allows the operator to develop an accurate mental model of the system and its behavior leading to calibrated trust in the system (Chen et al. 2014; Lee and See 2004).

Previous research has recommended that the system should make its purpose, process, performance (3Ps) and a history of 3Ps available to the operator to increase the operator’s understanding of the system (Lee and See 2004). Lee and See stated that both system capabilities and limitations should also be shown to the operator to assist in decision making. However, to reduce operator workload, this information should be in a simplified form to limit the amount of processing required for understanding and not overwhelm the operator (Cook and Smallman 2008; Neyedli et al. 2011). Thus, a transparent system should maximize operator decision-making performance and allow the operator to maintain overall situation awareness (SA) not only of the mission environment, but also of the state and intent of the system themselves (Chen et al. 2011; Endsley 1995).

The SA-based agent transparency model (SAT) (Chen et al. 2014) leveraged this effectiveness requirement and developed a useful theoretical framework to determine what type of information to display to an operator to maximize their situation awareness and assist the operator in developing an accurate mental model creating calibrated system trust. The SAT model builds upon the SA theory developed by Endsley (1995), the beliefs, desires, and intention agent framework (Rao and Georgeff, 1995), the 3P model (Lee and See 2004), and our previous work (Chen and Barnes 2012a; 2012b) (see Fig.1).

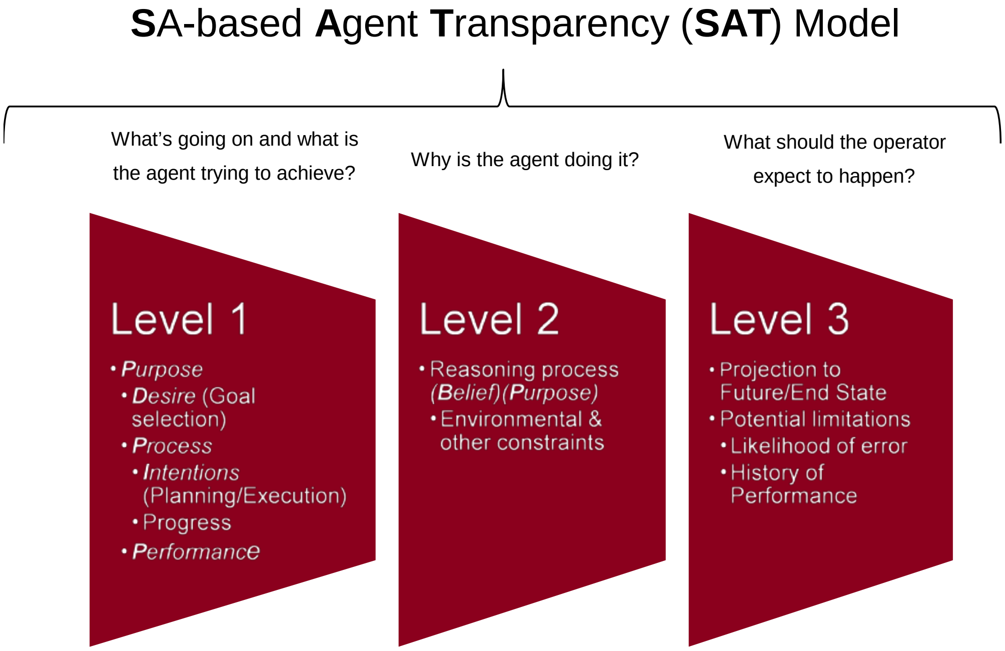

SA-based Agent Transparency (SAT) Model
What's going on and what is the agent trying to achieve?
Why is the agent doing it?
What should the operator expect to happen?
Level 1
• Purpose
• Desire (Goal selection)
• Process
• Intentions (Planning/Execution)
• Progress
• Performance
Level 2
• Reasoning process (Belief)(Purpose)
• Environmental & other constraints
Level 3
• Projection to Future/End State
• Potential limitations
• Likelihood of error
• History of Performance

Fig. 1 SA-based Agent Transparency model diagram (Chen et al. 2014)

In the first level, the operator is presented with basic information about the state of the world and the IA, such as the agent's current state and goals, intentions, and proposed actions. The second level builds connections between these basic pieces of rationale information to display the agent's current state and goals, intentions, and reasoning behind its proposed actions Finally, the third level provides the operator with information regarding the projection of future states of the system, such as the predicted consequences of the IA's decisions and any uncertainties associated with the systems actions (Chen et al. 2014). Previous research has supported the display of information that supports agent transparency to the operator as a way for mitigating uncertainties regarding a system's performance (Lyons and Havig 2014). Additionally, displaying a system's reliability, which has led to human operators adapting optimal reliance strategies (Wang et al. 2009), is similar to SAT Level 3, which suggests history of past performance can support optimal decision making (Chen et al. 2014). The benefits of including SAT-based information in an automated system are further supported by the notion that humans recalibrate their trust following automation failures when aware of system limitations (Dzindolet et al. 2003).

1.2 Trust in Automation

Proper calibration of trust is critical in high-risk situations, such as military operations (Groom and Nass 2007; Lee and See 2004). Over-reliance on an automated system when it is not appropriate (automation misuse) can lead to dangerous consequences, such as loss of life and property, while disusing the system when it can provide a benefit (automation disuse) is also erroneous, as the system could provide the operator with lower workload, faster response time, or greater performance the absence of which may be costly to the overall mission (Parasuraman and Riley 1997). Thus it is important for the operator to develop a properly calibrated trust in the system. Calibrated trust means that the operator has an accurate mental model of the system and relies on the system within the system's capabilities and is cognizant of its limitations, which leads the operator to override the system in situations outside of its limitations (Lee and See 2004).

Recent research suggested that the calibration of trust depends not only on the system's reliability but also on the perceived workload and usability (Hoff and Bashir 2015). In other words, displays that have more information to support transparency will be rated as more usable and more trustworthy because it is easier for the operator to form an accurate mental model of the system's 3Ps; however, more information does not always equate to relevant and good information. If the increased information processing requirements caused by the additional information shown to the operator increase workload, the display may be seen as less usable and will be trusted less. Therefore, we hypothesize that participant ratings of trust will increase linearly with increases in transparency level as the information displayed was developed based on the SAT model.

1.3 Workload

Another concern regarding autonomous systems is operator workload, which is the cost of performing a task that reduces an individual's ability to complete additional tasks (Cain 2007). Increased operator workload decreases performance and SA, and leads to incorrect automation usage decisions (Beck et al. 2007; Chen and Barnes 2012b; Parasuraman and Riley 1997). Operator workload is also a concern, as it may increase as agent transparency increases. Chen et al. (2014) stated that to support increased agent transparency, additional elements must be added to the interface; Lyons and Havig (2014) further stated that these additions may lead the operator to process more information, increasing workload. Conversely, the additional interface elements inform the operator of the current state, rationale, and future state projections so that the operator does not have to make these connections themselves, which may decrease their workload (Chen et al. 2011).

Consequently, the effect of increased agent transparency on workload is unclear. Thus workload will be an important factor in the current experiment. We hypothesize that workload will decrease with increased transparency level because the design of the information supporting agent transparency in the system is designed to lower operator cognitive load. However, we also note that increased workload is a valid concern when additional information is added to an interface and increased workload may decrease both trust in the system and perceived usability.

1.4 Usability

The International Organization for Standardization (ISO) defined usability as a user's effectiveness, efficiency, and satisfaction in a specific task context (Bevan 2009; ISO 2008). Previous research has found that greater usability was associated with more trust in automated systems (Wang et al. 2009) and calibrated trust while using automated decision aides (McBride and Morgan 2010). Further, displays that present information to support agent transparency may require integrating more and potentially complex information to operators, thus usability is a paramount concern when designing transparent autonomous systems (Beven 2009; Scholtz and Consolvo 2004). We hypothesize that usability scores will increase with transparency level because of` the previously hypothesized decrease in workload and increases in trust. In other words, the system will be perceived as more usable as it provides more information supporting of transparency to the operator. This hypothesis is also based on previous work indicating that workload decreases the effectiveness of a system (Beven and Macleod 1994).

1.5 Individual Differences

The effects of individual differences (IDs) on operator decision-making performance, workload, trust, and usability were evaluated in the present study. Several key individual differences were identified as relevant: perceived attentional control (PAC), spatial ability, working memory capacity (WMC), and gaming experience (GE).

1.5.1 Perceived Attentional Control

Attentional control refers to an individual's ability to self-regulate and enact effortful control over their attentional processes (Derryberry and Reed 2002). This ability assists individuals in determining which stimuli in the environment to direct their attention toward and assists in switching their attention between tasks (Astle and Scerif 2009). PAC is an individual's self-report of their ability to direct effortful control over their attentional processes (Derryberry and Reed 2002). Individual differences in PAC have been evaluated in previous studies involving supervisory control of multiple UxVs and may be an important predictor of performance in human-robotic interactions robot tasks (Chen and Barnes 2012b; Wright et al. 2013). Therefore, we hypothesize that individuals with increased PAC will be better able to calibrate their trust in the system by more quickly being able to determine issues with the IA. These individuals may also rate lower workload across all transparency levels.

1.5.2 Spatial Ability

Spatial abilities are another potentially important variable to consider explaining performance differences in supervisory multi-UxV systems. Previous research has found that individuals with greater spatial abilities not only made fewer performance errors on a robotic navigation task (Lathan and Tracey 2002) but also outperformed less spatially skilled individuals in threat detection tasks conducted while monitoring robotic performance (Chen and Barnes 2012a; Chen et al. 2008), in vehicle identification tasks (Fincannon et al. 2013), and during both direct line of sight and teleoperation navigation performance (Long 2011). We hypothesize that individuals with greater spatial abilities will specifically exhibit greater performance when given less information where spatial relations among assets is more valuable and may also exhibit lower workload during the experiment.

1.5.3 Working Memory Capacity

Greater WMC has been found to be associated with greater multitasking ability. High scores on the Operation Span (OSPAN), a measure of WMC, have been linked to greater performance on UxV tasks (de Visser et al. 2010). Further, when the demand for mental resources such as working memory capacity are overloaded, individuals must expend more effort and performance will decrease (Wickens 2008). We therefore hypothesize that individuals with greater WMC will self-report lower workload during the experiment and will have faster response times as a result of being able to more quickly and efficiently process the presented information.

1.5.4 Action Gaming Experience

Greater GE has previously been shown to increase accuracy and SA during multitasking situations (Chen and Barnes 2012b; Cummings et al. 2010). In fact, playing video games may assist individuals to develop strategies that can successfully be used to increase performance on other tasks. For example one study found that experienced action video game players (AVGPs), i.e., individuals who play action games such as first person shooters, outperformed nongamers in a change blindness task because they employed a broad search strategy. The nongamers on the other hand employed a more elaborate and costly strategy that cost additional time (Clark et al. 2011). We hypothesize, therefore, that action GE will be associated with faster reaction times throughout the experiment.

1.6 Current Study

In the current study we simulated a heterogeneous multi-UxV planning task where participants took on the role of an UxV operator whose job was to supervise vehicles and direct them to carry out missions while managing the commander’s intent plus vehicle and environmental constraints. Operators managed a team of 6 vehicles—2 unmanned aerial vehicles (UAVs), 2 unmanned ground vehicles (UGVs), and 2 unmanned surface vehicles (USVs)—to complete 3 blocks of 8 experimental events, 1 for each transparency level, for 24 discrete missions, using a simulator loosely based on the US Air Force Research

Laboratory's (AFRL's) FUSION system. The simulator uses a similar approach as the Playbook system developed by Miller et al. (2004) to engage the participant (called the operator) in mixed-initiative decision making. In the current experiment the simulator provides the operator with a particular play to call; however, it seeks input from the human by suggesting 2 similar plans (Plans A and B) to achieve the play. Each mission began with participants monitoring UxV vehicle positions and status. During this time they would receive 4 messages, each containing one of the following: patrol reports, updates on vehicle status, or commander intent messages. Two of the messages were relevant to the operators' task while 2 were irrelevant; messages were presented in a randomized order. Afterward, participants were given an objective to complete (e.g., locate a missing person or defuse an improvised explosive device (IED) along with 2 plans (Plans A and B) that may achieve that objective. The participants' task was to use information given to them by the IA to choose the best plan to complete each mission.

This experiment manipulated interface transparency level and either provided operators with a SAT Level 1 (basic plan information only), SAT Level 1+2 (basic plan information and reasoning), or SAT Level 1+2+3 (basic plan information, reasoning, and projections of uncertainty) interface. Our primary goal was to determine how increased information supporting of agent transparency based on the SAT model would affect operators' trust in the IA, workload, and their perceived usability of the system. Our secondary goal was to determine how individual differences affected the relationship between transparency and trust, as well as workload and usability. Finally, we are also interested in trying to understand participants' decision-making strategy and utilization of different elements of the interface to determine which parts of the display were useful to the participants.

2. Method

2.1 Participants

Thirty-five participants, recruited using an online participant pool, completed the experiment, and 5 were removed from the study. Two participants were removed due to technical issues, 2 because they did not pass the evaluation, and 1 because of a failed color vision test. Overall, 30 young adults in the Orlando, FL, area (18 men and 12 women) between the ages of 18 and 29 (*M* = 21.23, *SD* = 2.33) participated in this study. Participants were compensated $15/hr for their participation.

2.2 Apparatus

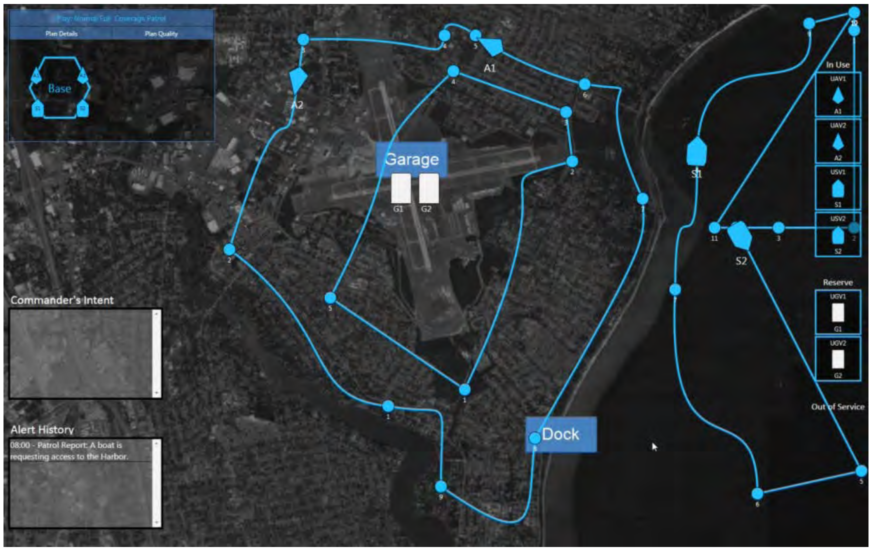
2.2.1 Simulator

A customized simulator, based on the AFRL's FUSION multi-UxV planning system (Spriggs et al. 2014), was created to support the current study. The simulator consisted of a standard desktop computer, one 60.96-cm (24-inch) monitor, one standard Windows keyboard, one standard 2-button mouse, 2 desktop speakers, and a customized software program. The simulator included several sections: a video window where participants watched UxV movements and received intelligence (Intel) messages, a mission assignment window where participants received a mission objective, and a decision window where participants received an asset capability tile to inform them of relevant information—vehicle capabilities, mission synopsis, Intel, and both of the IAs plan suggestions (Plans A and B).

Participants evaluated the 2 plans and selected the best plan based on their judgment. Additionally, participants were instructed to use 3 metrics to evaluate each plan: Speed, Coverage, and Capabilities. Speed was defined as how quickly each vehicle can arrive or carry out the mission. Coverage was defined as how well the vehicle can get "eyes on target" based on the type of sensors each vehicle carried. Finally, Capabilities was defined as the vehicle's appropriateness for the mission. Each vehicle had a set of strengths and weaknesses (e.g., can travel long distances, stealthy, or weaponized) that could affect Capabilities.

2.2.1.1 Simulator Video Window

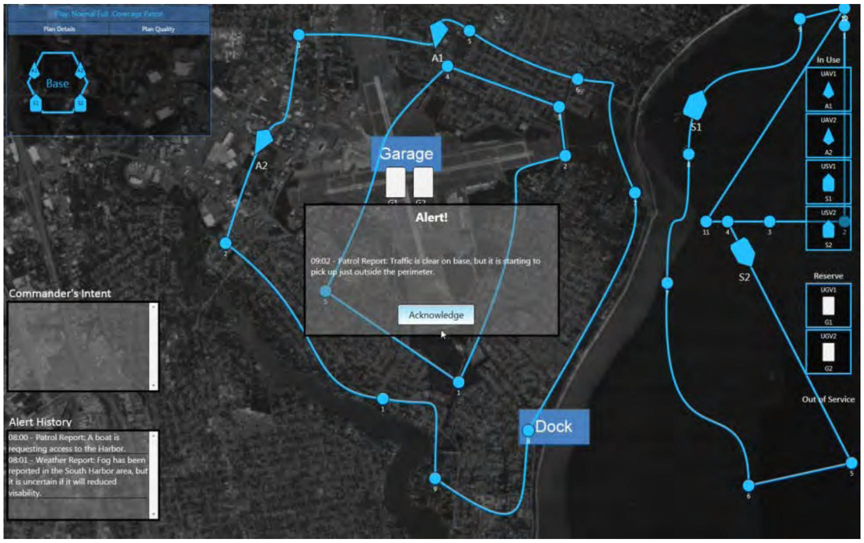
The simulator video window (Fig. 2) showed participants a base map, including each normal vehicle patrol route (interior road, perimeter, harbor, and sea lanes). The map was always displayed in the same orientation (north was always up). Locations of a garage and dock are denoted by labeled boxes. These locations house vehicles that are not currently assigned to the plan. UxVs were labeled using the middle letter of the appropriate vehicle acronym (A = UAV, G = UGV, S = USV) followed by a vehicle number (e.g., UAV1). Vehicle names remained constant during the experiment. Several smaller tiles were overlaid on the map, including a "play detail" tile (top left) that showed the play name and a visual representation of the current play, detailing active vehicle movement (colored to match the active play). In the video window, all vehicles began performing the "normal full coverage patrol base defense" play (always blue). An Intel history tile (bottom left) displayed previously acknowledged Intel. Messages from the base commander were prioritized and listed separately. Participants were given a scroll bar and could scroll if too much information was displayed in each box. A vehicle status tile (center right) identified which vehicles were in use (assigned to the active play), in reserve (unassigned to play), or out of service (grounded). Vehicles were displayed in one of 2 colors, the play's color or white (reserve or out of service). As participants watched the vehicles patrol the base, Intel messages would arrive (Fig. 3), which froze the simulation until they were acknowledged by clicking the Acknowledge box.



Alert History

08:00 - Patrol Report: A boat is requesting access to the Harbor.

Fig. 2 Simulator video window during opening video



Alert!

09:02 - Patrol Report: Traffic is clear on base, but it is starting to pick up just outside the perimeter.

Acknowledge

Alert History

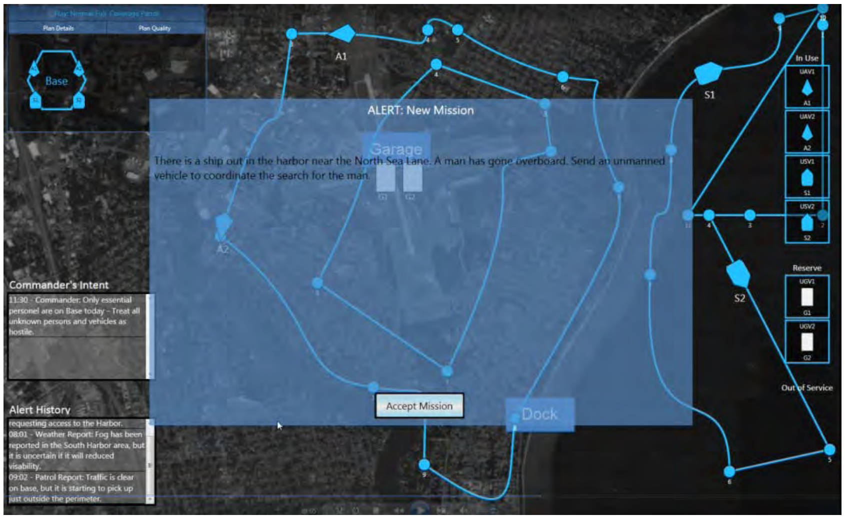
08:00 - Patrol Report: A boat is requesting access to the Harbor.

08:01 - Weather Report: Fog has been reported in the South Harbor area, but it is uncertain if it will reduced visability.

Fig. 3 Simulator video window with Intel message

2.2.1.2 Simulator Mission Assignment Window

The simulator mission assignment window (Fig. 4) was composed of an “alert” pop-up box that described the participants’ mission objective (e.g., “There is a ship out in the harbor near the North Sea Lane. A man has gone overboard. Send the best vehicle(s) to coordinate the search for the man.”) Participants then clicked on the “Accept Mission” box, which brought up the decision window.



ALERT: New Mission

There is a ship out in the harbor near the North Sea Lane. A man has gone overboard. Send an unmanned vehicle to coordinate the search for the man.

Commander's Intent

11:30 - Commander: Only essential personel are on Base today - Treat all unknown persons and vehicles as hostile.

Alert History

requesting access to the Harbor.

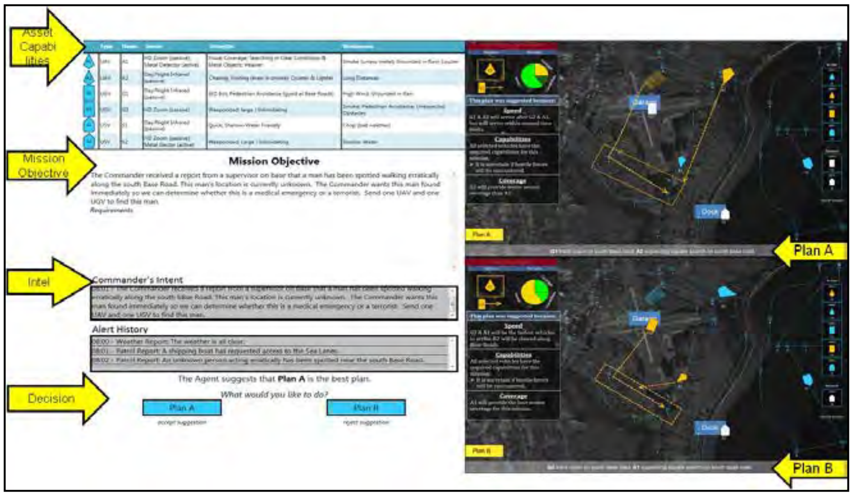
08:01 - Weather Report: Fog has been reported in the South Harbor area, but it is uncertain if it will reduced visability.

09:02 - Patrol Report: Traffic is clear on base, but it is starting to pick up just outside the perimeter.

Fig. 4 Simulator video window showing mission objective box

2.2.1.3 Simulator Decision Window

The simulator decision window (Fig. 5) displays the asset capabilities (top left), the mission objective (upper center left), the Intel messages (lower center left), the decision box (bottom left), an overview of Plan A (top right), and an overview of Plan B (bottom right).



Mission Objective

The Commander received a report from a supervisor on base that a man has been spotted walking erratically along the south Base Road. This man's location is currently unkown. The Commander wants this man found immediately so we can determine whether this is a medical emergency or a terrorist. Send one UAV and one UGV to find this man.

Requirements

Commander's Intent

08:01 - The Commander received a report from a supervisor on base that a man has been spotted walking erratically along the south Base Road. This man's location is currently unknown. The Commander wants this man found immediately so we can determine whether this is a medical emergency or a terrorist. Send one UAV and one UGV to find this man.

Alert History

08:00 - Weather Report: The weather is all clear.

08:01 - Patrol Report: A shipping boat has requested access to the Sea Lanes.

08:02 - Patrol Report: An unknown person acting erratically has been spotted near the south Base Road.

The Agent suggests that **Plan A** is the best plan.

What would you like to do?

Fig. 5 Simulator decision window

2.2.2 Eye Tracker

The SMI (SensoMotoric Instruments; Berlin, Germany) Remote Eye-tracking Device (SMI RED) was used to collect ocular indices to measure both visual attention and workload. The SMI RED system uses an infrared camera-based tracking system and allows for noncontact operation. The SMI RED uses a camera mounted under the computer monitor to track both the pupil and corneal reflection in both eyes. Eye movements were sampled at the rate of 60 Hz (each eye), which were logged in real time and synchronized with the simulator.

2.2.3 Survey and Tests

2.2.3.1 Demographics

A demographics questionnaire (Appendix A) was administered at the beginning of the study. This survey included information on participants' age, gender, education, computer experience, and GE. Participants rating of computer and GE was rated on a 6-point Likert-type scale (never, rarely, every few months, monthly, weekly, or daily). Frequent video gamers were categorized as individuals who reported playing either weekly or daily whereas nongamers were individuals who selected any other choice.

2.2.3.2 Color Vision Screening

Participants were given a screening for color deficiencies prior to participation using Ishihara color plates. Nine PowerPoint slides were shown to participants and only individuals who correctly answered at least 8 out of 9 were included in the study.

2.2.3.3 Trust Measures

We measured trust in 2 different ways in the current study. First, we measured participants' objective decision-making performance. Second, we measured participants' self-reported perceived trust in the IA, which is subjective in nature, using a questionnaire (Appendix B). Objective decision-making performance was measured by participants' reliance or rejection of the IA's prioritization of Plan A. The IA always presented Plan A as its indicator of the best option, and participants' were given a choice of accepting the IA's recommendation of Plan A (indicating trust in the agent) or choosing the alternative option Plan B, which indicated distrust in the IA. Participants made their decision by selecting one of the plan buttons displayed on the interface (Fig. 5).

Based on previous trust frameworks, an operator's appropriate reliance on the agent called calibrated trust (Hancock et al. 2011; Lee and See 2004) or appropriate learned trust (Hoff and Bashir 2015), the participant exhibits appropriate trust when the participant chooses Plan A when the IA is correct, and Plan B when it is not; in other words, the ideal state, from a signal detection theory perspective, is when a participant would only make hits and correct rejections. However, the participant may over-rely on the system and thus misuse it. In these situations, the participant would demonstrate high trust, but that trust would be associated with degraded performance (such as a high false alarm rate). Finally, participants may not trust the system and disuse it, even when Plan A is correct. In this situation, trust would also correlate with degraded performance (such as a higher miss rate; Table 1).

Table 1 Operational definitions of automation usage decisions used in the current study

Correct Plan	Automation	Operator	Usage	SDT[a]
A	A	A	Proper IA Use	Hit
B	A	B	Correct IA rejection	Correct rejection
A	A	B	IA disuse	Miss
B	A	A	IA misuse	False alarm

[a] SDT = signal detection theory.

Subjective trust was measured using the automation trust scale developed by Jian et al. (2000). However, this measure does not account for the types of automation suggested previously in the literature. For example, one seminal article (Parasuraman et al. 2000), proposed that each stage of information processing can be automated: 1) information acquisition (i.e., sensory processing), 2) information analysis (i.e., perception), 3) decision and action selection, and 4) action implementation (i.e., response selection). To this end, we combined the types of automation (Parasuraman et al. 2000) with the Jian et al. (2000) automation trust scale by asking each of the trust questions for each part of the information processing model. The current study, however, only manipulates the display of information already gathered (trust during information analysis) and performs decision and action selection. Consequently, we have only analyzed those scales and excluded information acquisition and action implementation from the current study. Each item was scored on a 7-item Likert-type scale (1 = not at all; 7 = extremely).

2.2.3.4 Response Time

Response time is defined as the time from when the decision window first appeared to the moment when the participant clicked one of the plan decision buttons and was measured directly from the simulation.

2.2.3.5 Workload Measures

We measured both objective and subjective workload. Objective workload was measured using eye tracking measures (e.g., fixation duration and pupil diameter). Subjective workload was measured using the National Air and Space Administration Task Load Index (NASA-TLX) (Hart and Staveland 1988; see Appendix C), a self-report questionnaire. The NASA-TLX measures a total weighted workload score based on 6 subscales: mental, physical, and temporal demands, as well as effort exerted, self-performance evaluation, and frustration felt during the task. Participants rated these 6 subscales on a continuous scale from 0 to 100, where lower scores on the scale indicate low workload and higher scores represent higher workload from that factor. Next, participants completed 15 pairwise comparisons (each scale appears 5 times) between each of the scale dimensions (e.g., effort vs. performance). Participants were instructed to pick the one factor that contributed more to their sense of workload during the task. To determine each subscale's weighting, the number of times each factor was chosen is divided by 15 (number of total comparisons).

Each subscale score is then multiplied by the weight to calculate the scale's weighted score. For example, if the Mental Demand scale was rated in the middle of the scale as a score of 50 and then chosen 5 of the 15 times during the pairwise comparisons, it would have a weighting of 0.33, which would be multiplied by its score of 50; thus, this factor would have a weighted score of 16.50.

2.2.3.6 Eye Tracking Measures

We measured 2 ocular indices of workload: fixation duration and pupil diameter. Both measures were captured only during the decision window to determine operator's workload during the decision process while they were interacting with the 3 transparency conditions. Both indices were averaged over the duration of each decision for each transparency level condition. Fixation duration, measured in milliseconds, is the time between saccades, when the eye is relatively still, during which visual information is processed (Holmqvist and Nyström 2011). Longer durations have been found to be associated with increased workload and cognitive processing (e.g., Yang et al. 2014). Pupil diameter, measured in millimeters, is the size of the pupil measured horizontally. Light levels were held constant throughout the experiment as changes in luminance can affect pupil size. Additionally, larger pupil diameters are associated with increased arousal and workload (Holmqvist et al. 2011).

2.2.3.7 System Usability Scale

The System Usability Scale (SUS) (Brooke 1996) is a 10-question scale designed to measure users' overall feelings of usability (efficiency, efficacy, and satisfaction) with the interface. The SUS is scored on a 5-point Likert scale (1 = strongly disagree; 5 = strongly agree) with half of the scale reverse-coded (Appendix D).

2.2.3.8 Attentional Control Survey

The attentional control survey (Derryberry and Reed 2002) is a 21-item survey scored on a 4-point Likert-type scale (almost never, sometimes, often, or always) to measure focused, selective, and divided attentional control (Appendix E).

2.2.3.9 Spatial Ability Measures

We used 3 measures of spatial ability. The first test was the Cube Comparison Test (Ekström et al. 1976; see Appendix F). This test displays sets of cube pairs and participants must mentally rotate the cubes to determine if they are the same cube from different orientations or different cubes. Second, the Spatial Orientation Test (Appendix G), modeled after the Cardinal Direction Test (Gugerty and Brooks 2004), evaluates participants' reorientation from an egocentric view. Participants view both a third-person view of a plane as well as a first-person view of a building. Participants have to determine the orientation of the building given the orientation of the plane. Finally, we used the Sense of Direction Scale (Kato and Takeuchi 2003; see Appendix H) which is a self-report 17-item survey, measured on a Likert scale, that measures 2 facets of spatial ability, memory for usual spatial behavior, and direction of orientation. These surveys measured 2 related but distinct components of spatial ability: spatial visualization (SpaV), which is the mental rotation of objects, and spatial orientation (SpaO), which is the reorientation of an environment (Hegarty and Waller 2004).

2.2.3.10 Working Memory Capacity

We used a version of the OSPAN task (Conway et al. 2005) to measure WMC. Participants alternated between solving a math problem in which they were instructed to press the space bar if the value of the equation equaled zero and being presented with a word. Sequence length was computer adaptive and increased with correct answers. After completing a sequence, participants were asked to recall the first letter of each word in the order of presentation. The average letters correct per sequence was used as our measure of WMC.

2.2.3.11 Personal Involvement

We created a novel measure of personal involvement (Appendix I) based on Zaichkowsky's (1985) Involvement with Advertisements Measure, which consisted of 6 questions scored on a 7-point Likert type scale (1 = not at all; 7 = extremely). Personal involvement in the task or task engagement was used as a potential covariate in the current study.

2.2.3.12 Structured Strategy Interview

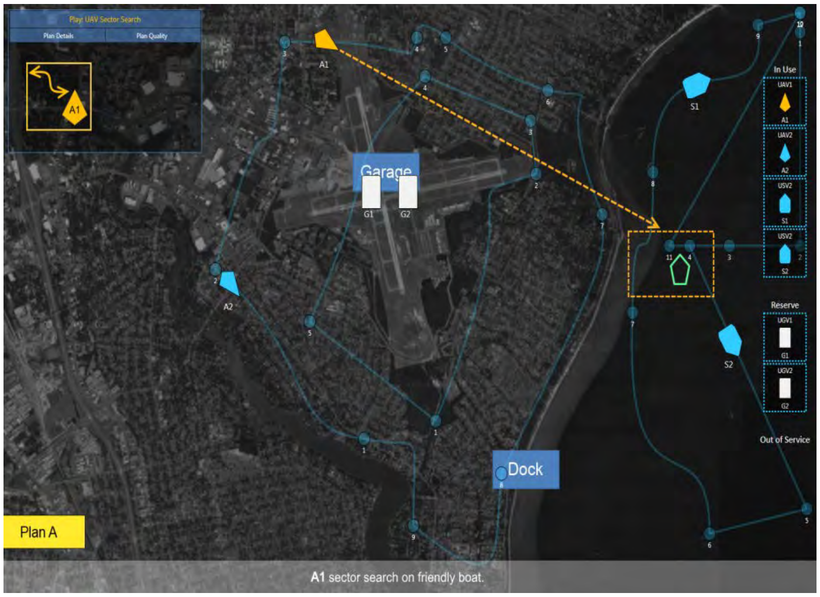
To determine participant's decision-making process after each block, we asked them to assign a numerical value to each element of the interface on a 7-point Likert-type scale (1 = not at all; 7 = extremely) and complete a series of qualitative, open-ended questions that asked them to describe the strategy they used during the previous block of trials (Appendix J). The experimenter took notes during the interview and completed a strategy sheet that was coded thematically.

2.3 Experimental Design

The experiment was a within-subjects design with 3 levels of IA transparency (SAT Level 1, SAT Level 1+2, and SAT Level 1+2+3) based on the SAT model (Chen et al. 2014). Transparency level was counterbalanced using a Latin square block design (Williams 1949). Participants completed 3 separate blocks, each consisting of 8 mission decisions of a single transparency level. The IA was incorrect 3 times, yielding a reliability rate of 62.5% based on Wickens and Dixon's (2007) finding that a 70% reliability rate is the point at which unreliable automation was worse than a lack of automation in terms of performance.

2.3.1 Transparency SAT Level 1

Transparency SAT Level 1 (Fig. 6) provided participants only with basic plan information. Participants were given the plan detail tile (top left), which displayed the play icon and play name, an informative bar at the bottom the screen that displayed a 1- to 2-sentence summary of the current plan, and the vehicle status tile (center right). Additionally, the map displayed the current status of the vehicles, their location and projected paths (represented as dashed lines), and areas of interest (e.g., targets, boats, and search areas).

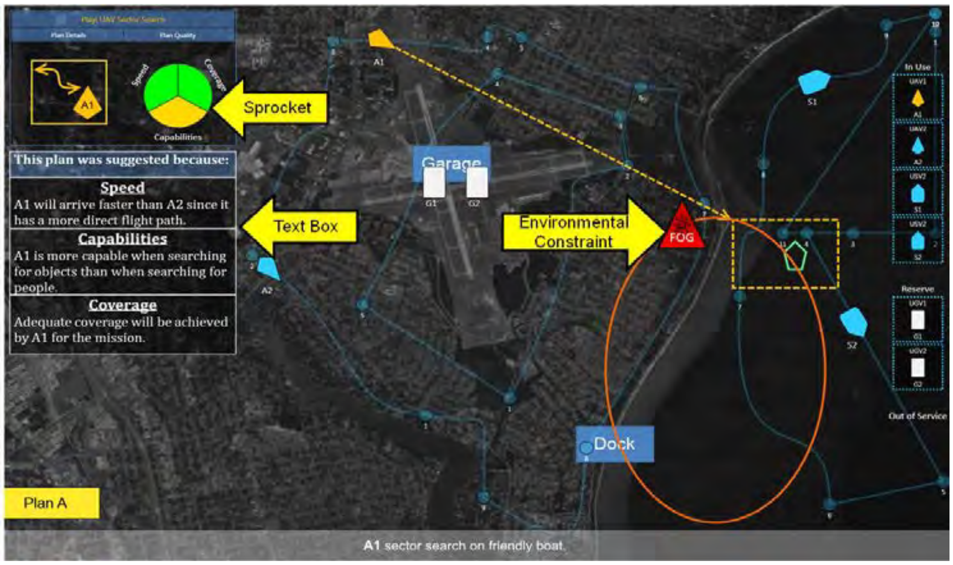


A1 sector search on friendly boat.

Fig. 6 Plan showing transparency SAT Level 1 condition

2.3.2 Transparency SAT Level 1+2

Transparency SAT Level 1+2 (Fig. 7) provided participants with all of the SAT Level 1 content and information regarding the IA's rationale. Participants were given the plan quality icon (sprocket); a text box describing factors that influenced the IA's recommendation of Plan A—speed, coverage, capabilities, environment alerts; and the IA's judgment of vehicle appropriateness to the mission.



This plan was suggested because:

Speed
A1 will arrive faster than A2 since it has a more direct flight path.

Capabilities
A1 is more capable when searching for objects than when searching for people.

Coverage
Adequate coverage will be achieved by A1 for the mission.

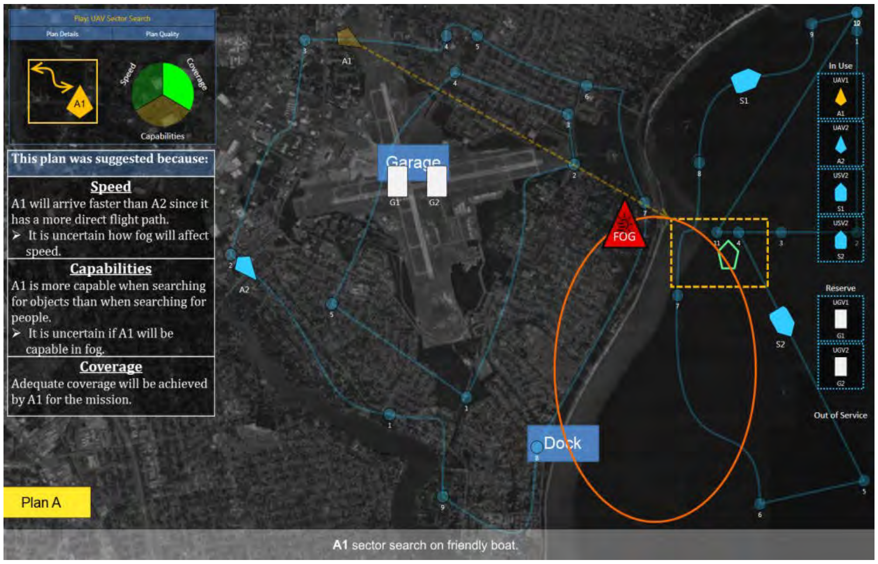
A1 sector search on friendly boat.

Fig. 7 Plan showing transparency SAT Level 1+2 condition. Labels indicate the location of the sprocket, text table, and a potential environmental constraint.

The sprocket had several parts, each displaying 2 types of information. First, the wedge size displayed the IA's judgment of the importance of each plan's evaluation metrics (larger wedge = higher importance). Second, each metric was colored either green (good) or yellow (average), based on the IA's determination of likelihood of mission success. A text box, displaying a written description of the plan's speed, coverage, or capability, was presented to participants underneath the sprocket. Speed was defined operationally as how quickly the vehicles can arrive to begin and complete the mission. Coverage was defined as the quality of sensor coverage provided during the mission. Capability is defined by specific strengths and weaknesses based on the specific equipment of each vehicle (displayed in the asset capability tile). Mission appropriateness of each vehicle was displayed to participants by manipulating vehicle icon size. In SAT Level 1+2, vehicle icons could be either smaller or larger. If larger, it was rated as most appropriate for the mission. Finally, environmental constraints that the IA used to consider its plan rationale were displayed on the map using a unique icon.

2.3.3 Transparency SAT Level 1+2+3

Transparency SAT Level 1+2+3 (Fig. 8) provided participants with all of the SAT Level 1+2 content and added projections of uncertainty to the interface. Three different types of projections of uncertainty were provided though the interface: 1) plan metric uncertainty (speed, coverage, and capabilities), shown as a transparent sprocket wedge and a bulleted statement in the text table, 2) vehicle uncertainty, shown as a transparent vehicle icon, and 3) route uncertainty, shown as a transparent vehicle route. Participants were not shown probabilities or likelihood comparisons; rather, just that the information was uncertain. Plan metric uncertainty was used to display a specific uncertainty about speed, coverage, or capability. For example, in Fig. 8, speed is green, meaning this metric is well satisfied by this plan; however, it is also uncertain. The specific reason is listed in the text box, as the current environmental condition may slow vehicle A2 down, reducing speed. The vehicle was uncertain because this condition may cause the vehicle to become less suitable for the mission, and the route was uncertain because it may have variability do to this same environmental factor.



This plan was suggested because:

Speed
A1 will arrive faster than A2 since it has a more direct flight path.
- It is uncertain how fog will affect speed.

Capabilities
A1 is more capable when searching for objects than when searching for people.
- It is uncertain if A1 will be capable in fog.

Coverage
Adequate coverage will be achieved by A1 for the mission.

A1 sector search on friendly boat.

Fig. 8 Plan showing transparency SAT Level 1+2+3 condition

2.4 Procedure

After participants completed the informed consent and were given a brief overview of the study, they completed a demographics questionnaire and a brief color-vision screening. Next, participants received experimenter-guided training that explained the tasks and knowledge needed to complete the study, including the interface. The training consisted of PowerPoint slides, 9 training missions (3 for each transparency level), and feedback performed using the simulator. The slides informed participants that the IA was not always 100% accurate but was reliable. The training session lasted approximately 45 min. After each training block, participants completed a brief structured interview about the strategy they used. Following training, participants received 18 evaluation missions. During the evaluation, participants were required to select 12 or more missions correctly to move onto the experimental missions. The evaluation lasted approximately 40 min. Participants were given a 5-min break, after which the eye tracker was calibrated, a process that consisted of the participants following a cursor around the screen using their eyes. Once the eye tracker was calibrated, participants moved on to the experimental missions.

Participants completed 3 blocks of experimental missions, one for each transparency level. Each mission was divided into 3 phases. The first phase consisted of participants viewing a team of UxVs (UGVs, UAVs, and USVs) as they patrolled the base perimeter for 45 s at the beginning of each mission. During this base defense task, participants received 4 Intel messages from either different patrols or the base commander, 2 of which had relevant Intel for the upcoming mission. Intel order was randomized. When an Intel message appeared, the simulation would pause to allow the participants to read and acknowledge each one individually. One message appeared every 9 s during the simulation, and the mission briefing appeared 9 s after the final Intel message at 45 s. During phase 2, occurring after the 45-s observation task, participants received a mission briefing with a specific objective. Participants were required to read and acknowledge this briefing. After acknowledging the mission objective, the participant entered the final, decision phase of the mission. The participant was presented with the decision window (Fig. 5), and the intelligence agent recommended 2 plans: Plan A (the agent's top choice) and Plan B (the agent's back-up choice). Participants chose between the 2 plans, and the next mission would begin. Participants completed 3 blocks of 8 events (1 for each transparency level), which were counterbalanced. Plan A was correct 5 times in each 8-mission block. After each block, participants completed the NASA-TLX trust survey, the personal involvement survey, the verbal strategy questionnaire, and the SUS. The experimental session lasted approximately 90 min, and the entire experiment lasted approximately 4 h.

3. Results

We present the results from a series of analyses of variance (ANOVAs) and multivariate analyses of variance (MANOVAs) across all dependent variables of interest: objective trust, subjective trust, response time, workload, and system usability. We also conducted a series of mixed ANOVAs on all of the individual differences variables for both objective trust and workload data across all transparency levels. We report results for SpaO, SpaV, WMC, and action GE conducted by completing a median split on each individual difference factor. Mixed ANOVAs were conducted because we used both a within subjects variable (transparency information) and between subjects variables for individual differences metrics (e.g., spatial ability and working memory).

All post hoc comparisons used a Bonferroni correction. Prior to all analyses, we screened for outliers and assumptions of multivariate normality with no significant deviations noted. We report effect sizes in terms of η^2 instead of partial η^2, as these can more easily be converted to R^2 and compared across studies (Levine and Hullett 2002).

3.1 Objective Trust

We report 2 separate analyses of objective trust. First, we conducted a signal detection theory analysis using the raw hit and false alarm data. Second, we calculated the proportion of IA proper usage rates and IA proper disuse rates to determine the effect of increasing transparency on calibrated trust.

3.1.1 Signal Detection Analysis

We used signal detection theory (SDT) to analyze participants' sensitivity to the IA's accuracy. We computed 2 indices of perceptual sensitivity from the hit (proper IA usage) and false alarm (IA misuses) data. The first index we used was the parametric index d' and the second index is the nonparametric $P(A)$, which is an estimate of the area under the Receiver Operating Characteristic curve described by a single hit and false alarm pair, also referred to as A' (Pollack and Norman 1964). The main advantage of using A' is that corrections do not have to be used in cases with hit rates of "1.0" or false alarm rates of "0" (e.g., Craig 1979; Davies and Parasuraman 1982); therefore, we report both metrics. For d' in cases of hit rates of "1.0" or false alarm rates of "0" we employed a correction to the data described by Macmillan and Creelman (2004), subtracting half of a hit and adding half of a false alarm to the data. In addition to perceptual sensitivity, we also calculated a measure of participants' response bias β.

3.1.1.1 Perceptual sensitivity (d′)

The results of a repeated-measures ANOVA on d' showed a significant transparency level effect, $F(2,58) = 11.39$, $p < 0.001$, $\eta^2 = 0.28$, where d' scores linearly increased with transparency information (Fig. 9). The greatest d' scores were found for transparency SAT Level 1+2+3 ($M = 2.75$, $SD = 1.01$), subsequently decreasing for SAT Level 1+2 ($M = 2.16$ $SD = 1.15$) and SAT Level 1 ($M = 1.55$, $SD = 1.05$). Post hoc tests using Bonferroni alpha adjustments within SPSS software (referred hereafter as post hoc tests) indicated a significant difference between SAT Level 1 and SAT Level 1+2+3 ($p < 0.001$) and a marginal difference between SAT Level 1 and SAT Level 1+2 ($p = 0.06$).

3.1.1.2 Perceptual sensitivity (A′)

The results of a repeated-measures ANOVA on A' showed a significant effect of transparency level, $F(2,58) = 7.54$, $p = 0.001$, $\eta^2 = 0.21$. A' scores linearly increased with transparency information. The greatest A' scores were found for SAT Level 1+2+3 ($M = 0.92$, $SD = 0.07$). Subsequently decreasing for transparency SAT Level 1+2 ($M = 0.87$, $SD = 0.14$) and SAT Level 1 ($M = 0.81$, $SD = 0.16$). Post hoc tests indicated a significant difference between SAT Level 1 and SAT Level 1+2+3 ($p = 0.002$). Fig. 9 shows both d' and A' across the 3 transparency levels. Means and standard error for both d' and A' are shown in Fig. 9, while means and standard deviations are shown in Table 2.

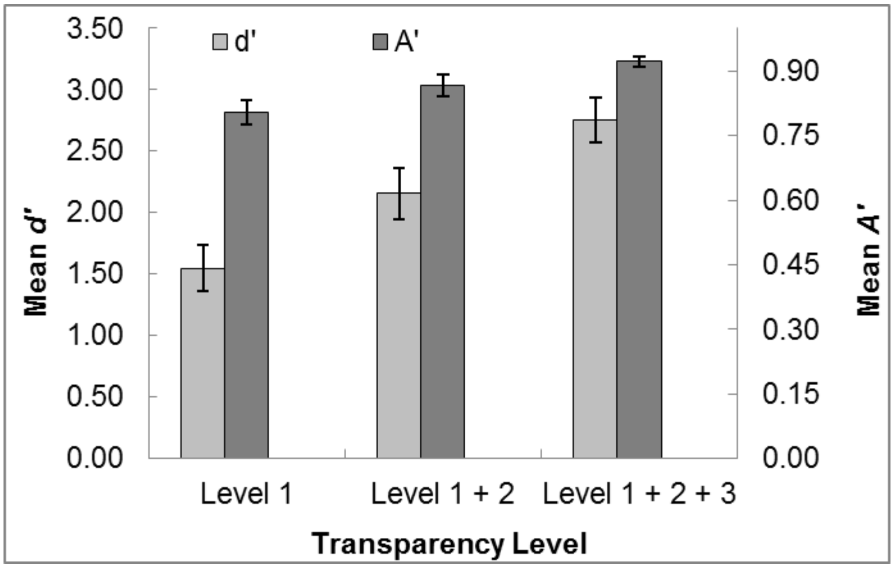


Fig. 9 Average *d'* and *A'* across transparency levels. Error bars indicate standard error of the mean (SEM).

Table 2 Perceptual sensitivity and response bias data across transparency levels

Dependent Variable	Transparency	*M (SD)*	95% CI[a]
	SAT Level 1	1.55 (1.05)	[1.16; 1.94]
Perceptual sensitivity (*d'*)	SAT Level 1+2	2.16 (1.15)	[1.73; 2.58]
	SAT Level 1+2+3	2.75 (1.01)	[2.37; 3.13]
	SAT Level 1	0.81 (0.16)	[0.75; 0.87]
Perceptual sensitivity (*A'*)	SAT Level 1+2	0.87 (0.14)	[0.82; 0.92]
	SAT Level 1+2+3	0.92 (0.07)	[0.90; 0.95]
	SAT Level 1	–0.23 (0.89)	[–0.10; 0.57]
Response bias (β)	SAT Level 1+2	0.68 (1.10)	[0.27; 1.09]
	SAT Level 1+2+3	0.06 (1.53)	[–0.51; 0.63]

[a] CI = confidence interval

3.1.1.3 Response bias (β)

Response bias was calculated using the likelihood ratio *β*. Results revealed no significant differences between participants' response bias in SAT Level 1 (*M* = –0.23, *SD* = 0.89), SAT Level 1+2 (*M* = 0.68, *SD* = 1.10), or SAT Level 1+2+3 (*M* = 0.06, *SD* = 1.53), *F* (2,58) = 2.51, *p* = 0.090, η^2 = 0.080. Overall, in our 3 transparency levels, participants were more likely to follow the IA's recommendation than to reject it, which may be due to our reliability manipulation. Response bias and sensitivity measures are shown in Table 2.

3.1.2 Proper IA Usage and Correct IA Rejection

Proper IA use and correction rejection rates represent a proportion of the 8 possible cases during each block. A repeated-measures MANOVA on both proper IA use and correct IA rejection rates across each transparency level was used to reduce pairwise error rate since both measures were moderately correlated (*r*'s = 0.26 – 0.73), but not so strongly correlated that warranted creating a composite measure. This analysis revealed a significant multivariate effect for transparency level using Wilks' Lambda criteria, F (4, 21) = 7.15, $p = 0.001$, $\eta^2 = 0.58$, $\Lambda = 0.42$. Since the multivariate effect was significant, we considered the univariate effects of each dependent variable separately.

3.1.2.1 Proper IA Usage Rates

Results for proper IA usage revealed a significant main effect of transparency level, F (2,58) = 12.33, $p < 0.001$, $\eta^2 = 0.30$. The greatest rate of proper IA usage was found in SAT Level 1+2+3 (M = 89%, SD = 12.15%), followed by SAT Level 1+2 (M = 87.44%, SD = 12.60%), while SAT Level 1 had the lowest proper IA usage rate (M = 75.85%, SD = 15.29%). Post hoc comparisons indicated participants' proper IA usage rates were significantly greater in SAT Level 1+2+3 ($p < 0.001$) and SAT Level 1+2 ($p = 0.003$) compared with SAT Level 1. There was no significant differences between proper IA usage rates between transparency SAT Level 1+2 and SAT Level 1+2+3 ($p = 1.00$).

3.1.2.2 Correct IA Rejection Rates

Results for correct IA rejection rates revealed a significant effect of transparency level, F (2,58) = 15.03, $p < 0.001$, $\eta^2 = 0.34$. The highest correct rejection rates were found in SAT Level 1+2+3 (M = 80.66%, SD = 19.97%), followed by SAT Level 1+2 (M = 67.11%, SD = 17.75%), while SAT Level 1 had the lowest correct rejection rates (M = 54.50%, SD = 20%). Post hoc comparisons indicated that participants' correct IA rejection rates were significantly greater in transparency SAT Level 1+2+3 than in SAT Level 1+2 ($p = 0.04$) and SAT Level 1 ($p < 0.001$). Furthermore, correct IA rejection rates in transparency SAT Level 1+2 were significantly greater than SAT Level 1 ($p = 0.013$). Results for both proper IA use and correct IA rejection rates for each of the transparency levels are displayed in Fig. 10.

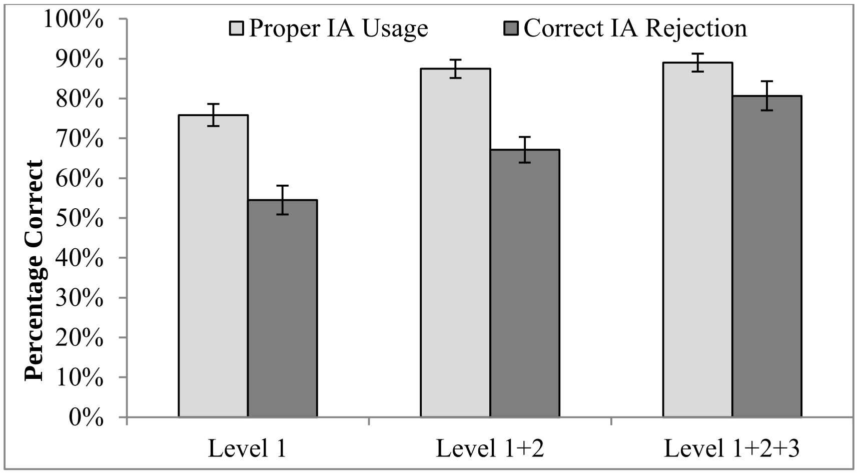


Fig. 10 Proper IA usage and correct IA rejection scores across transparency levels. Error bars indicate SEM.

Our next analysis was conducted on individual differences among groups. Our individual differences analyses revealed marginally significant interaction effect for working memory with a small effect, F (2,56) = 3.07, p = 0.054, η^2 = 0.01. The significant interaction effect was caused by performance differences in SAT Level 1 between low and high WMC groups (Fig. 11; Table 3). Individuals in the low WMC group had a lower proportion of correct rejections in SAT Level 1 (M = 0.47, SD = 0.18) than individuals with high WMC (M = 0.64, SD = 0.19; d = 0.92). While this pattern flipped in SAT Level 1+2 and SAT Level 1+2+3, these differences were small. Individuals in the low WMC group had a slightly greater proportion of correct rejections in SAT Level 1+2 (M = 0.68, SD = 0.18; d = 0.11) and SAT Level 1+2+3 (M = 0.82, SD = 0.19 d = 0.15) than individuals in the high WMC group SAT Level 1+2 (M = 0.66, SD = 0.18), SAT Level 1+2+3 (M = 0.79, SD = 0.22).

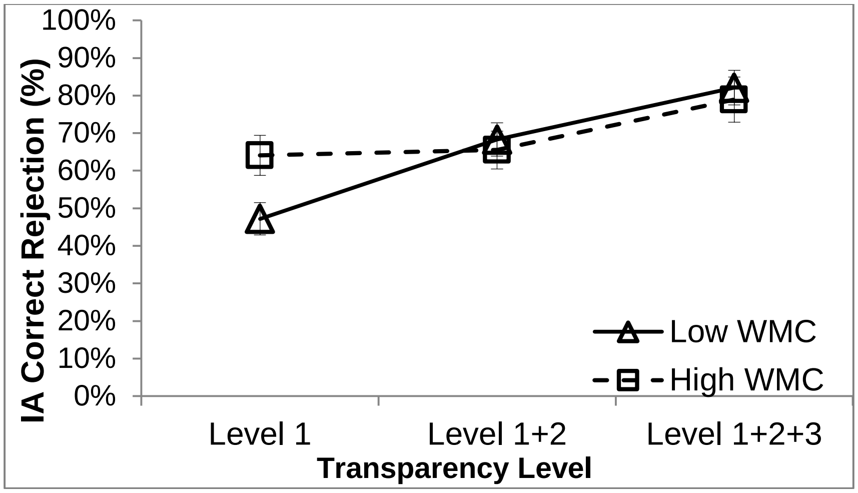


Fig. 11 Interaction of WMC on correct rejections across transparency level. Error bars indicate SEM.

Table 3 Individual difference (ID) factors for proper IA use (PU), correct IA rejection (CR), and response time (RT)

ID Factor	SAT Level 1			SAT Level 1+2			SAT Level 1+2+3		
	PU	CR	RT	PU	CR	RT	PU	CR	RT
Low SpaV	0.75 (0.20)	0.48 (0.20)	36.75 (18.00)	0.87 (0.13)	0.71 (0.20)	38.58 (21.95)	0.89 (0.14)	0.79 (0.20)	37.76 (22.63)
High SpaV	0.77 (0.09)	0.61 (0.19)	29.26 (15.04)	0.87 (0.13)	0.63 (0.14)	24.49 (11.42)	0.90 (0.11)	0.82 (0.20)	27.88 (12.06)
Low SpaO	0.74 (0.19)	0.51 (0.24)	30.28 (17.11)	0.88 (0.12)	0.68 (0.18)	28.91 (17.88)	0.90 (0.12)	0.80 (0.21)	28.21 (17.84)
High SpaO	0.78 (0.11)	0.58 (0.15)	35.73 (16.48)	0.87 (0.13)	0.66 (0.18)	34.16 (19.60)	0.88 (0.12)	0.81 (0.20)	37.43 (18.60)
Low WMC	0.74 (0.19)	0.47 (0.18)	32.57 (16.20)	0.89 (0.13)	0.68 (0.18)	33.26 (19.93)	0.92 (0.11)	0.82 (0.19)	32.43 (19.84)
High WMC	0.79 (0.09)	0.64 (0.19)	33.58 (18.06)	0.86 (0.12)	0.66 (0.18)	29.28 (17.29)	0.85 (0.13)	0.79 (0.22)	33.33 (17.41)
Non-AVGP	0.78 (0.17)	0.55 (0.21)	36.87 (17.89)	0.87 (0.13)	0.69 (0.18)	30.32 (16.43)	0.88 (0.13)	0.76 (0.19)	34.75 (19.00)
AVGP	0.73 (0.13)	0.54 (0.19)	27.20 (13.51)	0.88 (0.13)	0.64 (0.18)	33.35 (22.17)	0.90 (0.11)	0.87 (0.20)	29.92 (18.17)

Note: PU = proper IA use; CR = correct IA rejection; RT = response time; SpaV = spatial visualization; SpaO = spatial orientation; WMC = working memory capacity; AVGP = action video game player.

3.2 Subjective Trust

We conducted 2 separate between-subjects ANOVAs on both the information analysis and decision and action selection automation subscales. The need for between-subjects analyses stems from previous research, which has indicated that trust ratings can be biased based on prior experience with a system (e.g., Hoff and Bashir 2015), and thus we used the first block of trials that the participant experienced. Consequently, we analyzed trust for each subscale separately instead of creating a combined score.

There were no significant differences across transparency levels for the information analysis subscale, F (2,27) = 2.14, p = 0.14, η^2 = 0.14. Results did reveal a trend, where trust in the system's ability to integrate and display information increased as transparency level increased. Trust was greater in SAT Level 1+2+3 (M = 5.83, SD = 0.63), subsequently decreasing in transparency SAT Level 1+2 (M = 5.51, SD = 0.73) and SAT Level 1 (M = 5.19, SD = 0.70). Results for the "suggesting or making decisions" subscale were significant for transparency level, F (2,27) = 4.01, p = 0.03, η^2 = 0.23. Trust in the system's ability to suggest or make decisions increased as transparency level increased. Post hoc analysis revealed that trust was significantly greater in SAT Level 1+2+3 (M = 5.47, SD = 0.61) than SAT Level 1 (M = 4.63, SD = 0.88, p = 0.031). No significant differences were found between SAT Level 1+2 (M = 4.88, SD = 0.50) and SAT Level 1 (p = 1.0) or SAT Level 1+2+3 (p = 0.20), which are displayed in Table 4.

Table 4 Means, SD, and confidence intervals (CIs) for trust subscales across transparency levels

Dependent Variable	Transparency	*M* (*SD*)	95% CI
Information display and analysis trust subscale	SAT Level 1	5.19 (0.70)	[4.69; 5.69]
	SAT Level 1+2	5.51 (0.73)	[5.00; 6.03]
	SAT Level 1+2+3	5.83 (0.63)	[5.37; 6.28]
Decision action selection trust subscale	SAT Level 1	4.63 (0.88)	[4.00; 5.25]
	SAT Level 1+2	4.88 (0.50)	[4.52; 5.24]
	SAT Level 1+2+3	5.47 (0.61)	[5.03; 5.91]

3.3 Response Time

There were no significant difference in response time between all 3 transparency levels, F (2,58) = 0.38, p = 0.69, η^2 = 0.02 (Table 5).

Table 5 PU, CR, and RT data across transparency level

Dependent Variable	Transparency	*M* (*SD*)	95% CI
Proper IA use rate	SAT Level 1	0.76 (0.15)	[0.70; 0.82]
	SAT Level 1+2	0.87 (0.16)	[0.83; 0.92]
	SAT Level 1+2+3	0.89 (0.12)	[0.84; 0.94]
Correct IA rejection rate (%)	SAT Level 1	0.55 (0.20)	[0.47; 0.62]
	SAT Level 1+2	0.67 (0.18)	[0.60; 0.74]
	SAT Level 1+2+3	0.81 (0.20)	[0.73; 0.88]
Response time (s)	SAT Level 1	33.00 (16.73)	[26.76; 39.25]
	SAT Level 1+2	31.53 (18.63)	[24.58; 38.50]
	SAT Level 1+2+3	32.82 (18.51)	[25.91; 39.73]

ID analyses discovered a significant interaction effect for GE due to gamer differences in SAT Level 1 and SAT Level 1+2+3 (Fig. 12). AVGPs (*M* = 27.20, *SD* = 13.51) in SAT Level 1 had quicker response times than non-AVGPs (*M* = 36.87, *SD* = 17.89; *d* = 0.61). This pattern was also found in SAT Level 1+2+3 but to a lesser degree; the non-AVGPs (*M* = 34.75, *SD* = 19.00) had greater response times than AVGPs (*M* = 29.92, *SD* = 18.16; *d* = 0.26), *F* (2,56) = 5.74, *p* = .005, $\eta^2 = 0.17$.

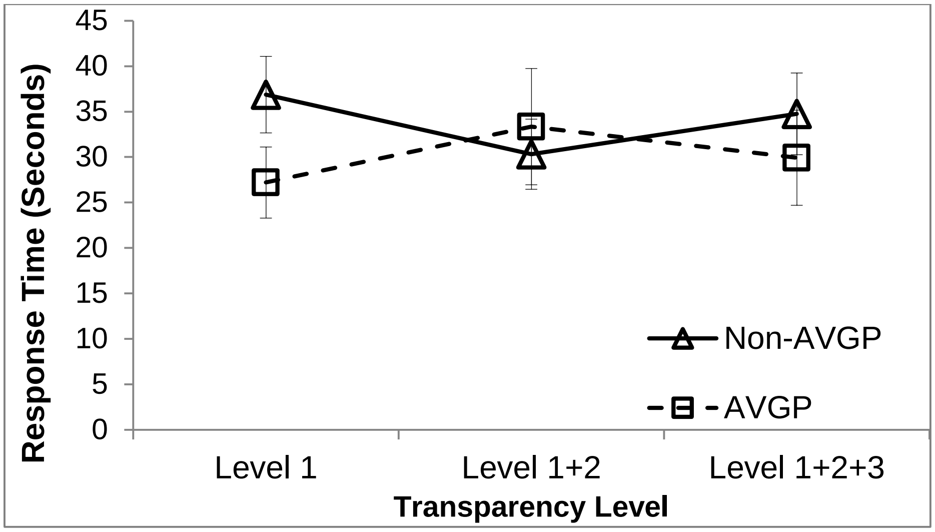


Fig. 12 Interaction of GE on RT across transparency level. Error bars indicate SEM.

3.4 Objective Workload

We report the results from our eye tracking analysis having removed 5 participants from the analysis pairwise across all conditions due to missing data, *n* = 25. In terms of objective trust, we did not find any effects of transparency level on either, fixation duration (FD), *F* (2,48) = 0.84, *p* = 0.44, $\eta^2 = 0.03$ or pupil diameter (PD), *F* (2,48) = 0.92, *p* = 0.91, $\eta^2 = 0.004$.

We found an interaction effect of SpaV on FD. Individuals with low SpaV had longer FDs in SAT Level 1 (*M* = 236.49, *SD* = 43.83) and 1+2 (*M* = 250.84, *SD* = 57.27) than individuals in the high SpaV group for SAT Level 1 (*M* = 218.42, *SD* = 42.19; *d* = 0.42) and SAT Level 1+2 (*M* = 222.78, *SD* = 38.84; *d* = 0.57). Interestingly, this pattern changed in SAT Level 1+2+3. The low SpaV group had shorter FDs in SAT Level 1+2+3 (*M* = 229.73, *SD* = 49.78) than those in the high SpaV group (*M* = 253.78, *SD* = 44.17; *d* = 0.51), *F* (2,46) = 6.19, *p* = 0.004, $\eta^2 = 0.20$ (Fig. 13 and Table 6).

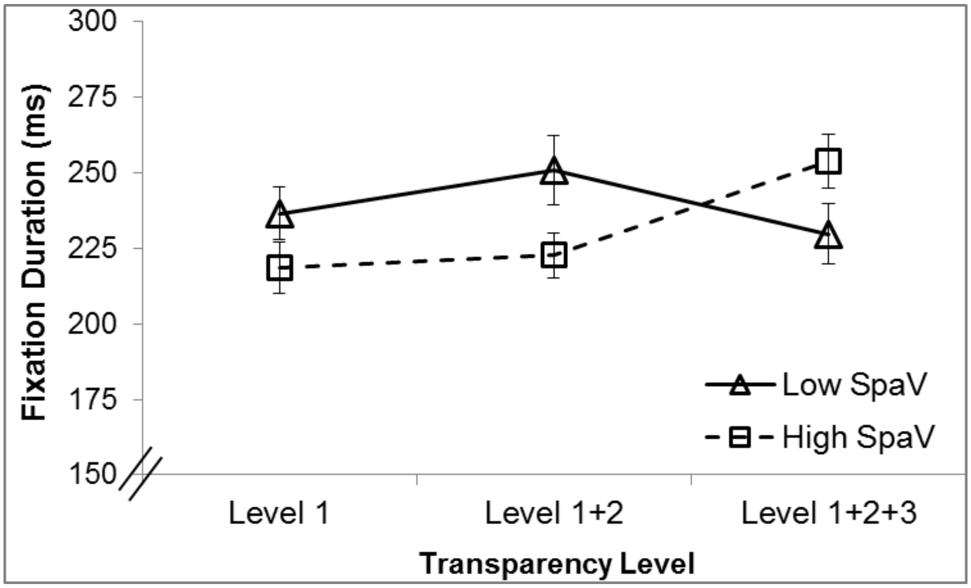


Fig. 13 Interaction effect of spatial ability on FD across transparency level. Error bars indicate SEM.

.

Table 6. Mean and standard deviations for individual difference factors across transparency level for eye tracking variables

ID Factor	SAT Level 1		SAT Level 1+2		SAT Level 1+2+3	
	FD	PD	FD	PD	FD	PD
Low SpaV	236.49 (43.83)	3.56 (0.52)	250.84 (57.27)	3.60 (0.51)	229.73 (49.78)	3.60 (0.67)
High SpaV	218.42 (42.19)	3.77 (0.36)	222.73 (36.84)	3.76 (0.35)	253.78 (44.17)	3.75 (0.36)
Low SpaO	223.39 (36.14)	3.47 (0.36)	233.15 (32.22)	3.53 (0.28)	235.32 (53.24)	3.50 (0.37)
High SpaO	233.45 (52.03)	3.90 (0.46)	242.70 (67.20)	3.87 (0.54)	248.85 (41.03)	3.90 (0.53)
Low WMC	216.46 (45.83)	3.68 (0.50)	229.37 (51.67)	3.71 (0.48)	231.98 (54.26)	3.68 (0.54)
High WMC	244.85 (34.04)	3.63 (0.40)	249.32 (46.49)	3.63 (0.40)	255.23 (33.92)	3.70 (0.40)
Non-AVGP	239.77 (38.00)	3.68 (0.50)	243.71 (47.79)	3.71 (0.49)	250.34 (47.91)	3.70 (0.55)
AVGP	206.56 (45.63)	3.63 (0.36)	226.04 (53.81)	3.61 (0.36)	225.16 (45.71)	3.61 (0.33)

Note: ID = individual difference, FD = average fixation duration, PD = average pupil diameter, SpaV = spatial visualization, SpaO = spatial orientation, WMC = working memory capacity, AVGP = action video game player.

We found a main effect of SpaO on PD. PD was larger for the high SpaO group (*M* = 3.89, *SD* = 0.17) than the low SpaO group (*M* = 3.50, *SD* = 0.03) across all transparency levels. The difference was larger in SAT Level 1 (*d* = 1.04) than SAT Level 1+2+3 (*d* = 0.96) or SAT Level 1+2 (*d* = 0.82), *F* (1,23) = 5.54, *p* = 0.027, $\eta^2 = 0.19$ (Fig. 14 and Table 6).

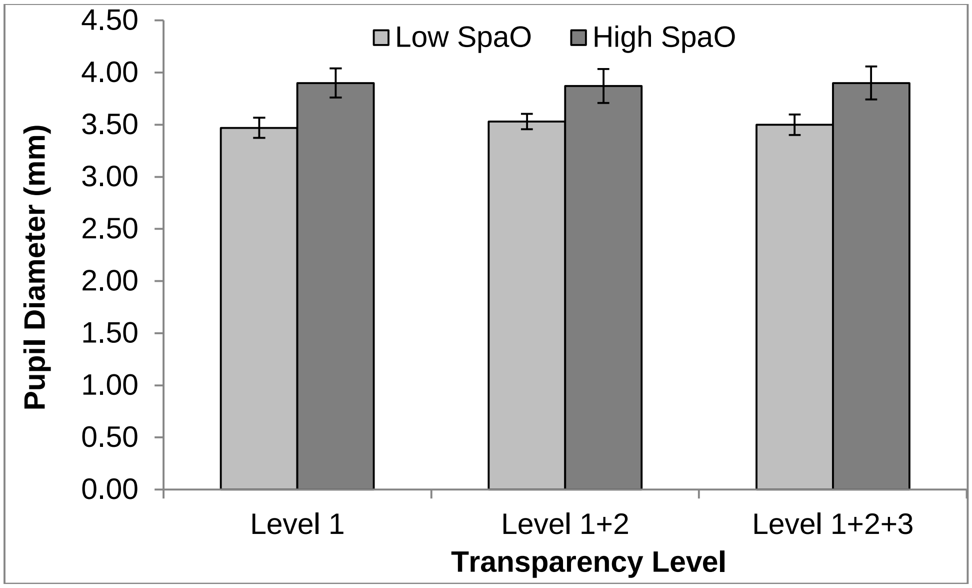


Fig. 14 Effect of spatial ability on PD across transparency level. Error bars indicate SEM.

3.5 Subjective Workload

We conducted a 6 (TLX subscale) × 3 (transparency level) repeated-measures MANOVA on the TLX subscales. The effect of the combined dependent variables was not significant using Wilks' Lambda criteria, *F* (12,18) = 1.14, *p* = 0.39, $\eta^2 = 0.43$, $\Lambda = 0.57$. In addition, no differences were found using the univariate ANOVAs among the individual subscales; therefore, we did not interpret these results. Fig. 15 shows each TLX subscale by transparency condition, and Table 7 displays means, standard deviations, and CIs across transparency levels for all subscales and global workload.

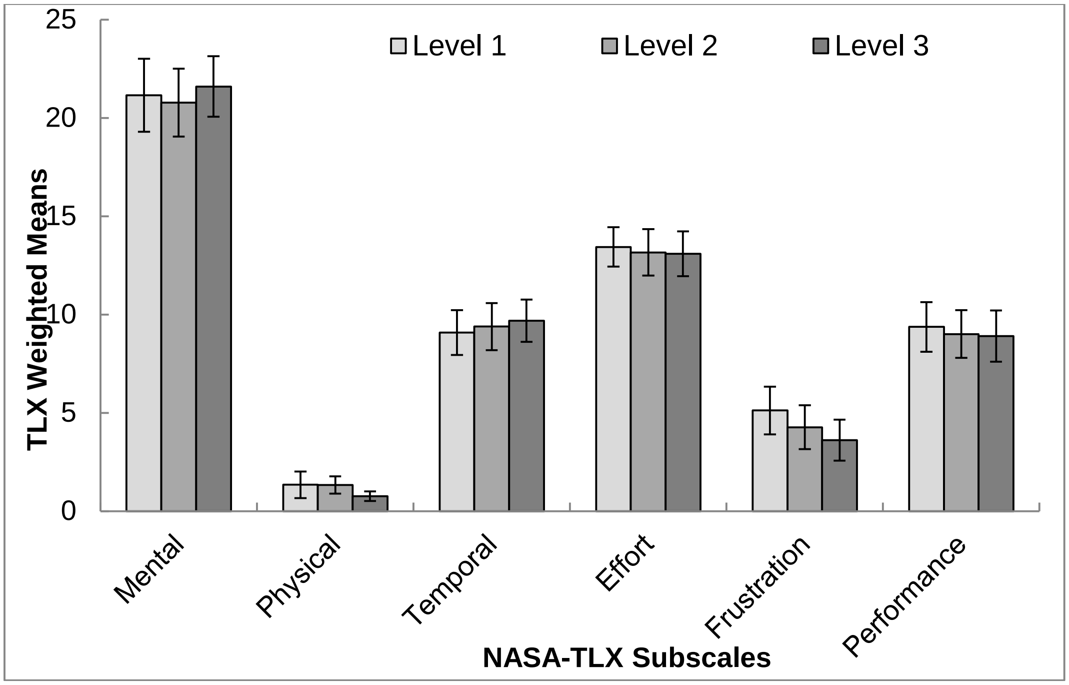


Fig. 15 Average weighted TLX subscale means across each transparency level. Higher numbers indicate greater workload, except for performance, where higher numbers indicate better perceived performance. Error bars are SEM.

Table 7 Means, standard deviations, and 95% CIs for subjective workload data across transparency level

Dependent Variable	**Transparency Level**					
	SAT Level 1		**SAT Level 1+2**		**SAT Level 1+2+3**	
	***M*(*SD*)**	**95% CI**	***M*(*SD*)**	**95% CI**	***M*(*SD*)**	**95% CI**
Mental	21.16 (10.17)	[17.25; 24.95]	20.78 (9.46)	[17.25; 24.31]	21.60 (8.45)	[18.45; 24.76]
Physical	1.34 (3.71)	[–0.04; 2.73]	1.33 (2.41)	[0.23; 2.03]	0.76 (1.38)	[0.25; 1.28]
Temporal	9.08 (6.26)	[6.74; 11.42]	9.39 (6.57)	[6.94; 11.84]	9.69 (5.87)	[7.50; 11.88]
Effort	13.44 (5.84)	[10.74; 15.57]	13.16 (6.46)	[10.74; 15.57]	13.09 (6.27)	[10.74; 15.43]
Frustration	5.12 (6.63)	[2.65; 7.60]	4.27 (6.14)	[1.97; 6.56]	3.61 (5.69)	[1.49; 5.74]
Performance	9.37 (6.93)	[6.78; 11.95]	9.01 (6.68)	[6.52; 11.51]	8.90 (7.14)	[6.24; 11.57]
Total workload	59.51 (17.85)	[52.85; 66.18]	57.73 (18.25)	[50.92; 64.55]	57.66 (17.25)	[51.22; 64.10]

Because we did not find any statistically significant differences between conditions, we conducted an analysis by collapsing across the 3 transparency level conditions to determine if significant differences existed between the TLX subscales for the experimental task as a whole using an ANOVA. The analysis revealed significant differences among the individual TLX subscales, F (5,29) = 43.55, p < 0.001, η^2 = 0.56. Mental demand (M = 21.17, SD = 8.65) was the greatest overall contributor of workload (all comparisons p < 0.001). In addition, the effort subscale (M = 13.23, SD = 5.20) was the next greatest contributor of workload, which was greater than physical workload (M = 1.08, SD = 2.39, p < 0.001) and frustration (M = 4.33, SD = 5.57, p < 0.001).

3.6 Usability

We conducted a repeated-measures ANOVA on system usability scale total scores. The analysis revealed a significant effect for transparency level, F (2,58) = 5.70, p = 0.006, η^2 = 0.11. Post hoc comparisons indicated that participants found the system more usable in both transparency SAT Level 1+2+3 (M = 66.75, SD = 19.40, p = 0.02) and SAT Level 1+2 (M = 66.42, SD = 18.61, p = 0.07) than in SAT Level 1 (M = 61.83, SD = 22.77). No significant differences were found between SAT Level 1+2 and SAT Level 1+2+3 (p = 1.00) (Table 8).

Table 8 Means, SD, and CIs for SUS data across transparency levels

Dependent Variable	Transparency	M (SD)	95% CI
SUS total score	SAT Level 1	61.83 (20.77)	[54.08; 69.59]
	SAT Level 1+2	66.42 (18.61)	[59.47; 73.37]
	SAT Level 1+2+3	66.75 (19.40)	[59.51; 74.00]

3.7 Personal Involvement

We conducted a repeated-measures ANOVA on personal involvement scores to determine if involvement varied across transparency levels. The analysis did not find any significant differences across transparency level, F (2,58) = 1.43, p = 0.247, η^2 = 0.047 (Table 9).

Table 9 Means, SD, and CIs for personal involvement data across transparency levels

Dependent Variable	Transparency	M (SD)	95% CI
Personal involvement score	SAT Level 1	30.40 (5.14)	28.48; 32.32
	SAT Level 1+2	30.73 (4.74)	28.96; 32.50
	SAT Level 1+2+3	31.43 (4.64)	29.70; 33.16

3.8 Decision-making Strategy

We divided these data into 2 separate analyses. Quantitative data were subjected to a series of ANOVAs, while the qualitative data derived from the structured interviews were coded into themes and analyzed descriptively.

3.8.1 Quantitative Data

We conducted a series of repeated-measures MANOVAs to determine if differences existed between the participant's perceived usefulness of different display elements within each level during the experiment.

3.8.1.1 SAT Level 1 information

Participants received SAT Level 1 information in every block of the experiment. They received 1) the name of the current play (Play Name), 2) detailed maps of the plan to complete the play (Play Details), 3) color-coded plans to indicate which vehicles were included in the plan (Plan Colors), 4) the status of each vehicle (Vehicle Status), 5) a brief summary of each plan (Information Bar), and 6) an experimental aide designed to reduce participant workload by providing them the strengths and weaknesses of each UxV (asset capability tile). Therefore, we analyzed the results in a 6 (display elements) × 3 (transparency level) MANOVA. The analysis did not reveal any differences between display element use rates of SAT Level 1 information between transparency levels using Wilks' Lambda criteria, F (10,20) = 1.11, p = 0.39, η^2 = 0.36, Λ = 0.64. However, we observed that the mean for the asset capability tile appeared to vary across transparency levels, and the MANOVA may have concealed differences for the use of the asset capability tile. Therefore, we conducted a separate univariate ANOVA for the asset capability tile to reveal any differences hidden by the MANOVA, and the analysis showed a significant difference, F (2,58) = 4.17, p = 0.20, η^2 = 0.13. Therefore, we believe that the other nonsignificant findings masked the differences for the asset capability tile (Fig. 16). The asset capability tile was perceived as significantly more useful in transparency SAT Level 1 (M = 6.10; SD = 1.2; p = 0.006) than in transparency Levels 2 (M = 5.50; SD = 1.5) or 3 (M = 5.47; SD = 1.5; p = 0.25).

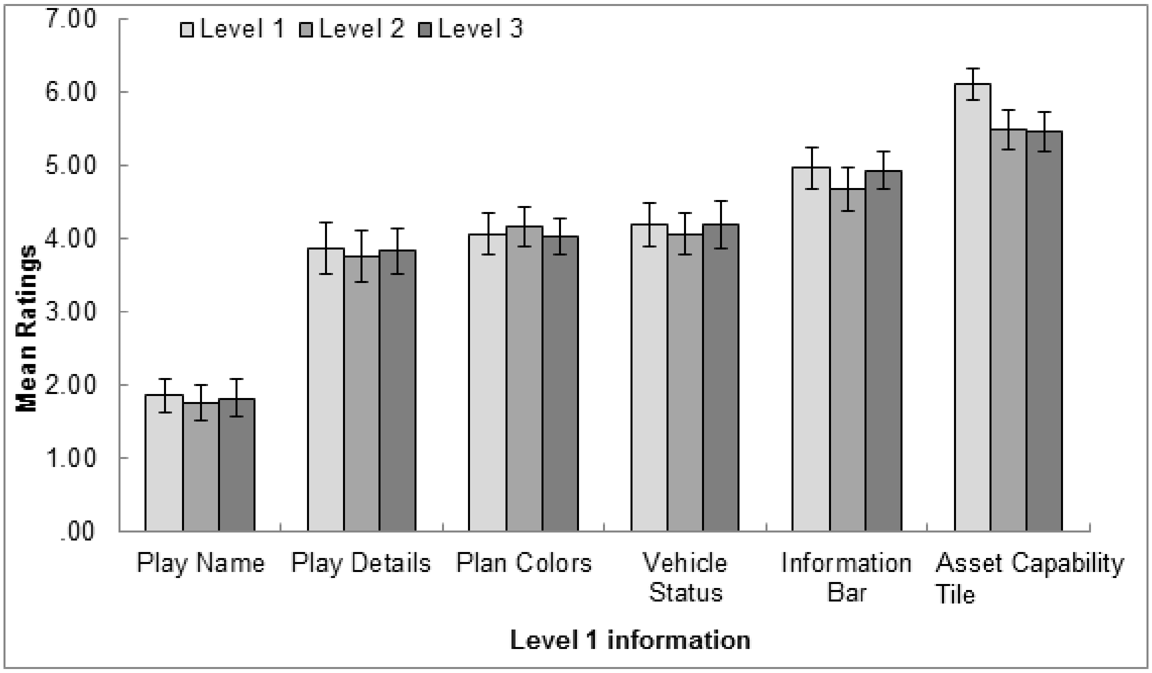


Fig. 16 Usefulness ratings across transparency level conditions specifically for Level 1 user interface elements

In addition, usefulness ratings were assessed within each specific level using a repeated-measures ANOVA to determine which specific display elements participants found most useful to completing their decisions. Within transparency SAT Level 1 specifically, a main effect of information type was found, *F* (5,145) = 19.65, $p < 0.001$, $\eta^2 = 0.40$. The play name was, significantly, the least helpful piece of rationale information given to the participant by the system (*M* = 1.87, *SD* = 1.28, $p < 0.001$), while the asset capability tile was found to be the most helpful piece of rationale information given by the system in SAT Level 1 (*M* = 6.10, *SD* = 1.90, $p < 0.001$). The remaining display elements did not differ significantly from each other.

3.8.1.2 SAT Level 2 information

Participants received SAT Level 2 information only in 2 conditions (transparency Levels 2 and 3; Fig. 7) of the experiment. Participants received 6 pieces of SAT Level 2 rationale information: 1) the size of the vehicle to indicate UxV capabilities (Vehicle Size); 2) environmental overlays on the maps (e.g., wind, fog, roadway debris; icons on map [alerts]); 3) the sprocket, which was divided into overall as well as 4) Wedge Size and 5) Wedge Color, and 6) IA reasoning information provided in a table for each plan (Table Text). Therefore, we analyzed the results in a 6 (display elements) × 2 (transparency level) MANOVA. The analysis did not reveal any differences between information use rates of SAT Level 2 information between transparency levels using Wilks' Lambda criteria, *F* (4,25) = 0.69, $p = 0.61$, $\eta^2 = 0.10$, $\Lambda = 0.91$. A comparison of the means revealed no significant difference between SAT Level 2 and SAT Level 3 and almost identical values between ratings of transparency SAT Level 2 and 3; therefore, SAT Level 2 information was used similarly across both transparency levels (Fig. 17).

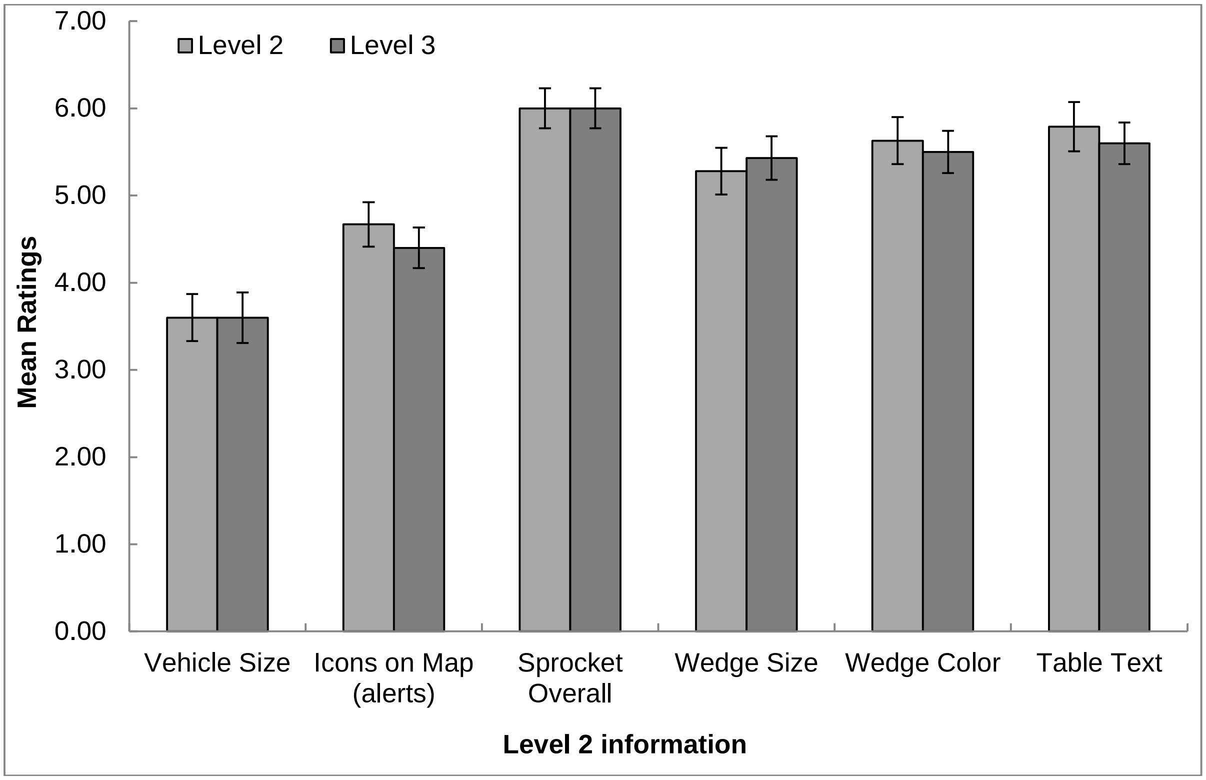


Fig. 17 Usefulness ratings across transparency level conditions specifically for SAT Level 2 user interface elements

Additionally, usefulness ratings were assessed within each specific level, using repeated-measures ANOVA, to determine which specific display elements participants found most useful to completing their decisions. Within transparency SAT Level 2 specifically, a main effect of information type was found, *F* (5,145 = 9.66, $p < 0.001$, $\eta^2 = 0.25$. The sprocket (*M* = 5.96, *SD* = 1.29, $p < 0.001$) and the text table (*M* =5.75, *SD* = 1.53, $p < 0.001$) were perceived as the most useful display elements given to the participant by the system. Within the sprocket, the wedge color (*M* = 5.62, *SD* = 1.50) was seen as significantly more helpful than the wedge size (*M* = 5.28, *SD* = 1.44), *t* (29) = 2.28, *p* = 0.03, Cohen's *d* = 0.23. The remaining display elements did not differ significantly from each other.

3.8.1.3 SAT Level 3 information

Participants received SAT Level 3 information only in the transparency SAT Level 3 block of the experiment and we were primarily concerned with the transparency of the sprocket and the text table (Fig. 8); therefore, we analyzed the results using a paired t-test. The analysis did not reveal any differences between the participants' information use rates of the sprocket displaying uncertainty compared with the table displaying uncertainty *t* (29) = 1.14, p = 0.27, Cohen's *d* = –0.20, indicating that no differences exist between the usefulness of these pieces of uncertainty (Fig. 18).

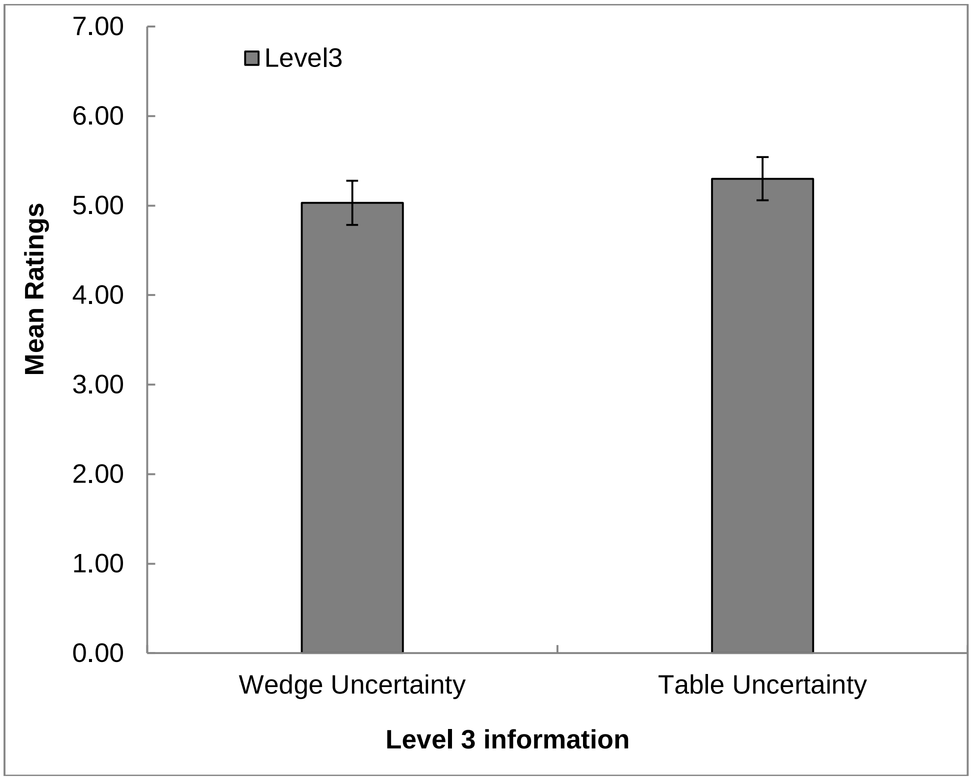


Fig. 18 Usefulness ratings across transparency level conditions specifically for SAT Level 3 user interface elements

3.8.2 Qualitative Data: Differences between Training and Experiment

We also used information gleaned from structured interviews to determine if there was a strategy change between training and the experimental sessions. During training, participants may use many strategies and through experience may change those strategies based on the feedback given. It was expected that during the experimental blocks participants would discontinue strategies that were problematic during training given that only those who do well on the evaluation phase of the experiment continue to this phase. We used the qualitative data to support this analysis by looking at the number of strategies and the change between strategies from training to experimental sessions. Answers to open-ended questions on the strategy interview were grouped into strategies and analyzed by testing the number of strategies between training and experimental blocks across all transparency levels during the experiment.

3.8.2.1 SAT Level 1 Strategies

A repeated-measures MANOVA was conducted to determine the differences between training and experiment on SAT Level 1 information, and no differences were found as a main effect, F (5,25) = 1.35, p = 0.28, η^2 = 0.21, Λ = 0.78, or as an interaction between transparency level and session, F (10,110) = 1.00, p = 0.41, η^2 = 0.08, Λ = 0.84. We reported partial eta squared as reported in SPSS for our MANOVA results η^2 (Fig. 19).

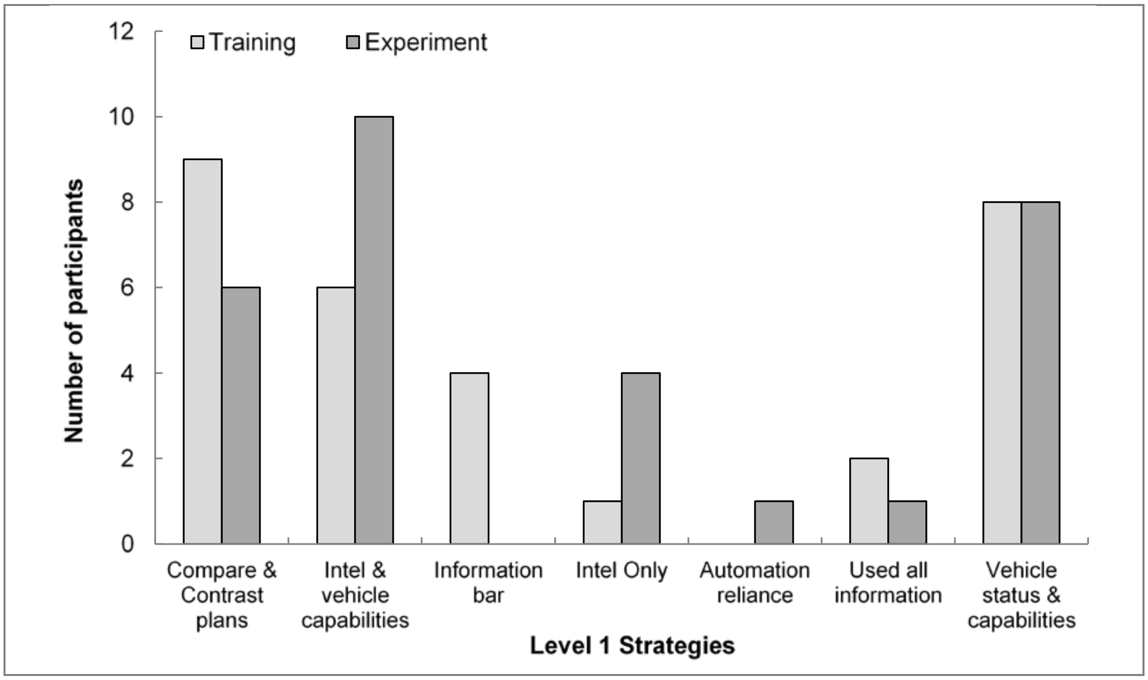


Fig. 19 Number of participants using specific strategies during both training and experiment for the transparency SAT Level 1 condition

The qualitative data indicate that while participants perceived no differences in the usefulness between the different parts of the interface, they used that information in different ways. For SAT Level 1, participants tried to use the differences between the 2 plans during training, while in the experimental block they used Intel much more as they progressed through training and into the experimental blocks.

3.8.2.2 SAT Level 2 Strategies

A repeated-measures MANOVA was conducted to determine the differences between training and experiment on SAT Level 2 information, and no differences were found as a main effect, F (4,25) = 1.12, p = 0.38, η^2 = 0.15, Λ = 0.85, or as an interaction between transparency level and session, F (4,25) = 0.82, p = 0.41, η^2 = 0.06, Λ = 0.94 (Fig. 20).

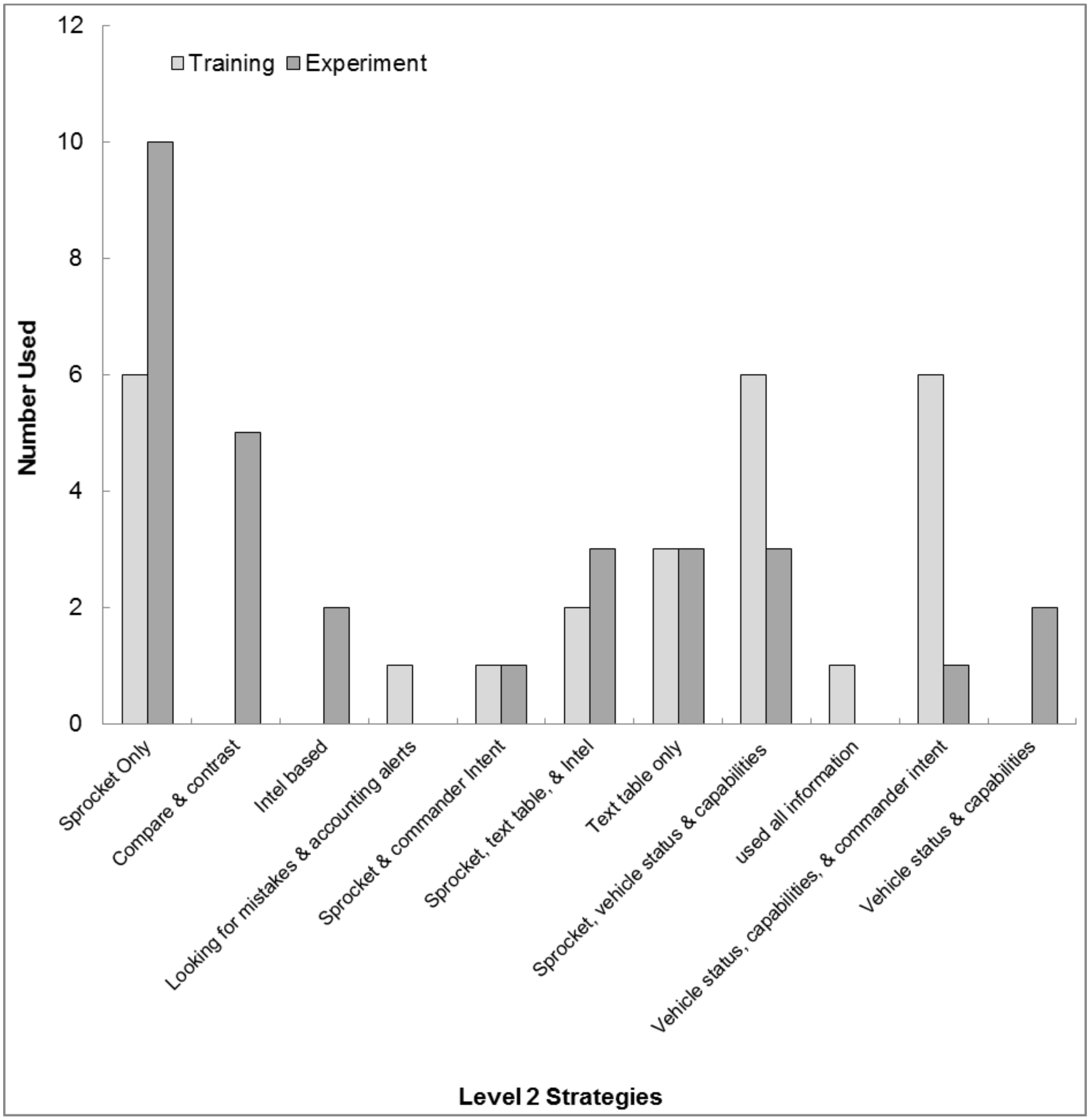


Fig. 20 Number of participants using specific strategies during both training and experiment for the transparency SAT Level 1+2 condition

In SAT Level 2, during training, participants used the sprocket and tried to understand capabilities, while during experiment they used the Intel, compared the plans, and then used the sprocket to finalize their decision, while others only used the sprocket.

3.8.2.3 SAT Level 3 Strategy

A repeated-measures MANOVA was conducted to determine the differences between training and experiment on SAT Level 3 information, and no differences were found as a main effect, F (2, 28) = 1.56, p = 0.23, η^2 = 0.10, Λ = 0.90 (Fig. 21).

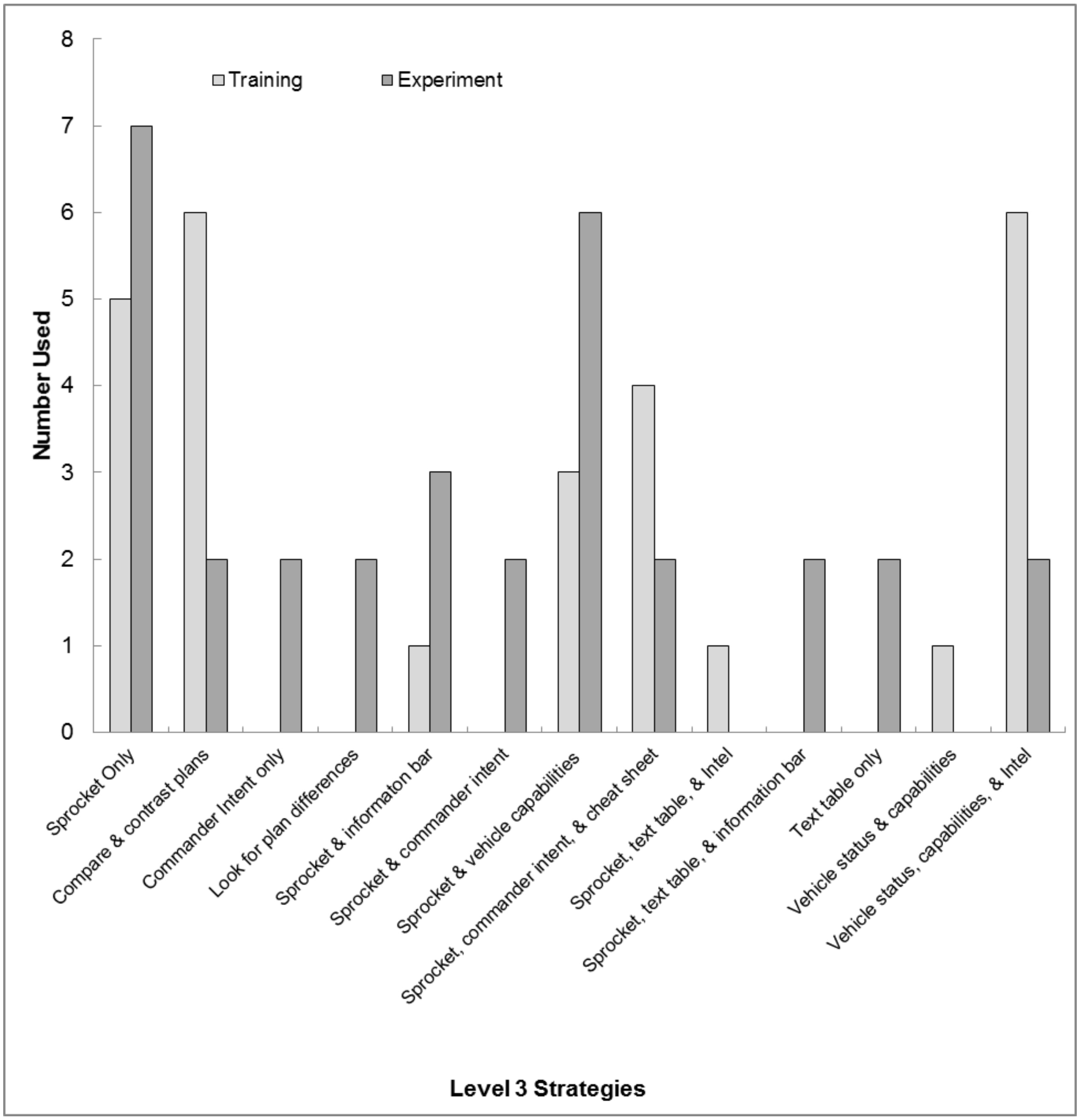


Fig. 21 Number of participants using specific strategies during both training and experiment for the transparency SAT Level 1+2+3 condition

The strategies used in SAT Level 3 were similar to the strategies used in SAT Level 2 except many more people tried to compare the plans during training, while during the experiment they relied much more heavily on the sprocket and starting using the table due to the additional information that was provided.

4. Conclusions and Discussion

The current study had several goals. The primary objective was to study the effect of transparency level on user trust and to determine any potential performance trade-offs that may occur with respect to response time or workload. Across all mission events, participants were assisted by an IA that presented them with 2 plans with which to achieve the mission objectives and satisfy commander's intent. Participants had to weigh vehicle capabilities, locations, and Intel with the IA's assessment of plan success, potential uncertainties, and recommendations to achieve mission success. The IA made optimal recommendations 62.50% of the time; failure to do so was due to information that had not yet been processed. Thus, the primary measure of trust was the participant's automation usage decision to accept or reject the system's recommendation of Plan A. Secondary goals of the study included an exploration of the effects of an operator's perception of system usability and the implications of individual differences (spatial ability, WMC, and action GE) on trust, workload, and ratings of the utility of each display element used in the experiment.

The objective trust data supported our hypothesis that increases in information to support operator transparency lead to greater IA with proper uses and correct IA rejections in Level 1+2 and even more in Level 1+2+3, which means both disuse and misuse decisions decreased across transparency levels as well. The addition of reasoning information in Level 1+2 increased proper IA use by 11% and correct rejection rate by 12%. The addition of both reasoning and uncertainty information in Level 1+2+3 improved proper IA use by an additional 2% and correct rejection rate by an additional 14%. Taken together, the proper IA usage and correct IA rejection percentages indicate that objective trust calibration increased linearly as a function of transparency level, with Level 1+2+3 proper IA usage rate at 90% and correct IA rejection rate at 81%. This increase suggests that incorporating information regarding both reasoning and uncertainty into heterogeneous tactical decision making successfully allowed our participants to indicate a more accurate trust in the IA shown by more accurate performance when making tactical decisions. These results are consistent with Helldin et al.'s (2014) findings that information that supported increased transparency also increased task performance as well as with Finger and Bisantz's (2002) findings that displaying uncertainty information can support operator decision making. This relationship between performance and agent transparency was also supported by our SDT analysis. Level 1+2+3 yielded participants with both the highest d' and A' values, indicating the greatest level of perceptual sensitivity and paralleling the findings of the objective trust data.

Using automation usage decisions as an objective measure of trust, however, only partially gauges operators' trust in the IA. Participants may not have trusted the system at all, disregarded the IAs recommendation, and manually solved each mission. This would indicate that our analysis of our objective trust data was flawed, as overall system disuse should decrease when it becomes more transparent. Therefore, subjective trust measures are needed to provide further insight into the participants' trust. Subjective trust results, using the modified Jian et al. (2000) scale, aligned with our objective trust findings; therefore, we reject that operators failed to trust the IA and that the more-parsimonious explanation that the increased transparency based on the SAT model increased operator trust in the agent. Results for the "integrating and displaying information" and the "decision and action selection automation" subscales provided evidence that our operators trusted the IA's recommendation more when the system was more transparent. This result is consistent with the findings of Oduor and Wiebe (2008), in which transparency positively affected trust calibration, suggesting that trust in human-agent teams is an important factor in performance (Freedy et al. 2007). Further, we mention calibrated trust because we hypothesized that calibrated trust would be associated with both greater objective and subjective trust in the IA. Additionally, subjective trust subscale ratings were sensitive to our system reliability manipulation. We manipulated the reliability of the IA's recommendation, but the information supporting agent transparency in the current system provided to the participants was accurate for each mission. Therefore, trust was greater for information analysis automation, which was accurate 100% of the time compared with decision and action selection, which was accurate only 62.5% of the time for all transparency levels. Taken together, these findings suggest that participants displayed appropriate trust calibration.

The individual differences analyses failed to find significant individual differences between either of our spatial ability or perceived attention control measures with objective or subjective trust scores across transparency levels. Previous studies have found that differences in both perceived attentional control and spatial abilities were key to understanding differences in task performance while managing multi-robotic systems (Chen and Barnes 2012a, 2012b; Chen et al. 2008; Lathan and Tracey 2002). Therefore, we hypothesized that perceived attentional control and spatial abilities would be important factors in our task. However, our study differed in several key aspects from previous supervisory control studies, which may have lessened individuals' use of attentional or spatial skills. Previous human-robot interaction studies have all used sensor feeds from cameras as a component of their performance or decision-making tasks. In these tasks, participants either had to use visual information to teleoperate an unmanned vehicle (Chen et al. 2008; Lathan and Tracey 2002) or complete a threat detection task while making route decisions for a team of robots (Chen and Barnes 2012b). Performance on our task required integrating information displayed by both the IA and Intel to determine if the IA was basing its decisions on a faulty premise or incomplete information. Further, our tasks did not specifically require manipulating objects or the robot's location in space and the map always remained in the same orientation; thus, the operators' spatial abilities were used less frequently than in the aforementioned teleoperation studies.

We also looked at another key individual difference, WMC, with regard to objective trust. We found that individuals with low WMC had worse performances in Levels 1 and 1+2+3 while WMC did not significantly vary in Level 2. Participants were given only basic information in Level 1, and, to make accurate decisions, they had to process and synthesize the information given to them; therefore, presumably, individuals with higher WMC performed better at this task due to that capacity. Interestingly, we found a similar pattern for Level 1+2+3, indicating that while participants were shown uncertainty information, they also had to determine what the effects of uncertainty are within the context of each specific mission. Previous research has indicated that uncertainty information adds working memory load and, consequently, individuals with low WMC are more likely to rely on heuristics to resolve their memory load than individuals with higher WMC (Quayle and Ball 2000).

Another potential effect of adding additional user interface elements to support agent transparency is that individuals stop using basic information for reasoning or uncertainty elements (i.e., sprocket or text table). The results of the analyses performed on strategy differences between transparency level conditions indicated that all Level 1 elements, except for the asset capability tile, were used similarly across transparency levels. The asset capability tile was not a specific user interface element but rather an experimental addition to prevent novice participants from having to memorize asset capabilities and sensor payload information; therefore, the finding that individuals rated the asset capability tile as more helpful in the Level 1 condition serves as a manipulation check. As the agent became more transparent, users did not have to do as much work to determine the correctness of the IA's decisions, and thus the asset capability tile became less useful during Level 1+2 and Level 1+2+3 conditions. The condition order was counterbalanced, so we can reject any potential confounds from participant usefulness ratings and experience using the asset capability tile information. We also found that participants rated Level 2 user interface elements as similarly helpful across both level 1+2 and Level 1+2+3 conditions. Overall, the sprocket and text table were rated as the most helpful pieces of Level 2 information. The ratings for Level 3 information found no significant differences between the uncertainty displayed in the sprocket or the text table. This finding was similar to that of the Level 2 information where the sprocket was also rated as more helpful than the text table. The structured strategy interviews revealed that, typically, participants primarily relied on the sprocket and used the table information as a secondary source of information. This finding is logical because the sprocket was a very salient element in the display that conveyed information about priority, reasoning, and potential uncertainties presented in the plan.

Our analysis between different overall strategies used between training and experimental blocks indicated several differences as operators learned to complete the task. During training, individuals are brand new to the system and have not had much experience using the interface. During the experimental blocks, they were much more experienced since participants had already gone through training as well as the evaluation blocks, indicating that the change from training to experiment denotes stopping strategies that may not be particular useful and increased use of strategies that participants found helpful. During the Level 1 condition, participants placed emphasis on comparing and contrasting the differences between plans and looking at the information bar. During the experimental missions, they found these strategies less useful and instead focused on both capability differences between the assets and using the Intel received during each mission. During the Level 1+2 condition, participants initially used vehicle capability, Intel, and text table; however, after training, participants placed a greater emphasis on the sprocket, comparing the differences between the plans using the sprocket with many participants only using the sprocket as their sole strategy. This same pattern was also found in Level 1+2+3. Overall, this finding indicates that participants spent less time manually checking each plan and instead relying more on the system displays (sprocket and text table) and using the Level 1 information to confirm or double-check their trust in the system.

With regard to our analyses of the response bias data, we did not find significant differences that would indicate a particular bias between the IA's recommended and backup plan for all transparency levels. Overall, decisions were somewhat liberal, which was expected due to the greater percentage of Plan A scores. This finding further suggests that complacency did not appear to contribute to the participant's decision in the current study. One reason for this finding may have been the lack of workload differences between transparency levels. Greater levels of workload may have forced a certain level of reliance on the system due to the cost of manually solving each decision. Previous studies have found significant workload differences between different levels of automation assistance (Wright et al. 2013). Additionally, other domains found that increased workload can negatively affect a decision-making performance, causing operators to sacrifice performance rather than optimize performance (Cummings 2006). In the current experiment, complacency could have occurred during SAT Level 1 due to the decreased transparency of the system and the level of task-related information provided to the participants (Parasuraman et al. 1993). At SAT Level 1+2 and SAT Level 1+2+3, we hypothesized complacency could occur due to the increased amount of information in the display that may create higher mental demand (Inagaki and Itoh 2013). Therefore, the nonsignificant response bias findings indicate that operator responses were a function of the level of information provided by the system, rather than an indicator of workload. As previously stated, we failed to find any significant differences in workload across conditions measured objectively or subjectively.

In addition, we also failed to find increased response time as a function of transparency level. Previous research has found speed accuracy trade-offs as well as an association between speed and workload indicative of additional processing requirements (Helldin et al. 2014). Our results indicated that participants' perception of mental demand and effort were consistently greater than other subscales across transparency levels, suggesting that the task primarily stressed mental demand and effort that is appropriate for the task supporting the validity of the subjective workload findings. Additionally, our analysis of the TLX data revealed an interesting potential trend. Frustration decreased as transparency information increased. This finding, along with our other findings, align well with the SUS findings, which indicated that participants' perceived usability of the system increased as transparency information increased. Thus, usability could be an underlying factor in participants' performance, trust, and workload.

We further analyzed eye-tracking data (fixation duration and pupil diameter) as a measure of objective workload. Aligning with the TLX scores, we did not find a change in workload as we added transparency information. The goal of transparency information is to mitigate workload by offloading information synthesis to the IA and displaying that information to the operator in a meaningful way. These findings indicate that the benefits of transparency may not introduce potential costs for workload supporting this idea instead of increased costs for implementing transparency.

While no significant differences in workload were found generally, we accessed the effects of our individual differences measures. In doing so we found an interesting dissociation between spatial abilities and our eye tracking measures. We found significant differences between high and low SpaV groups for fixation duration but not pupil diameter. This dissociation occurred between SpaO operators as well, with significant differences for pupil diameter but not fixation duration. Fixation duration has been linked to workload and stress as well as more effortful scene processing. Some studies have shown shorter fixation durations for higher workload conditions that lead to greater visual scanning (Van Orden, et al. 2001). Therefore, it appears participants with lower spatial abilities were under more stress than those with higher spatial abilities. Individuals with higher spatial abilities were able to focus on more critical parts of the interface while those with lower spatial visualization skills were forced to scan around the interface.

Decreased pupil diameter has previously been found to be an indicator of fatigue (Holmqvist et al. 2011); thus, it is possible that individuals with lower spatial orientation skills became fatigued faster by the amount of information in the interface than those with greater spatial abilities. Increased pupil diameter, on the other hand, has also been associated with increased cognitive workload (Van Orden et al. 2001). This possibility is seemingly at odds with the previous findings, as both SpaV and SpaO are related abilities. Since individuals with higher SpaO are better able to integrate and process information from the environment and take different perspectives given a single egoentric viewpoint, we believe that those with higher spatial abilities tried to directly compare the maps among each other, while individuals with lower spatial abilities may not have relied on other information in the environment. This possible behavior may also explain the interaction in

SAT Level 1+2+3 as individuals with higher SpaV abiliites may have attempted to integrate the projections of uncertainty into the map to better understand its effects to the plan.

4.1 Future Work

In the current experiment we defined our SAT Level 3 manipulation as the display of uncertainty. While uncertainty is a key variable that affects how conclusions are made on projected information, after further consideration we believe that uncertainty may apply to each level of the SAT model, as it does not solely reflect the system's future state. For example, sensor errors may make basic information uncertain, and uncertainties in SAT Level 2 may make the operator circumspect of the IA's reasoning process. Additionally, both of these uncertainties can be separated from SAT Level 3 uncertainties about the IA's projections of future states. Therefore, future research should investigate how incorporating uncertainty into each level of transparency could affect performance, trust, and workload. We have one such study planned using a more ecologically valid interface, the AFRL Fusion test bed; however, information used to support agent transparency may be somewhat contextually dependent, meaning each system needs to determine which display elements are best for that specific system.

4.2 Conclusion

Our findings are increasingly important to facilitate decision making between the human operator and complex automated systems. Since automation is a key part of future systems, operators will need to rely on advanced automation, such as IAs, to enhance mission effectiveness due to the increased level of information flow (Paas and Merriënboer 1994). We examined the level of information received from the IA needed to create an effective, transparency interface, specifically addressing 3 issues: performance, trust, and workload.

Unlike Helldin et al. (2014), who found that increased transparency resulted in increased performance and trust calibration at the cost of greater workload and longer response time, our results support the addition of transparency information, loosely based on the SAT model (Chen et al. 2014). The addition of transparency information greatly improved decision-making accuracy and perceptual sensitivity without cost to speed or increased workload. These findings align well with our trust and usability data. Both trust subscales suggest that participants trusted the IA's recommendation more when the system was more transparent. Similarly, SUS findings indicated that participants' perceived usability of the system increased as transparency information increased.

5. References

Astle DE, Scerif G. Using developmental cognitive neuroscience to study behavioral and attentional control. Developmental Psychobiology. 2009;51(2):107–118.

Beck HP, Dzindolet MT, Pierce LG. Automation usage decisions: controlling intent and appraisal errors in a target detection task. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2007;49(3):429–437.

Bevan N. Extending quality in use to provide a framework for usability measurement. In: Kurosu M, editor. Human centered design. Heidelberg (Germany): Springer; 2009. p. 13–22.

Bevan N, Macleod M. Usability measurement in context. Behaviour and Information Technology. 1994;13(1–2):132–145.

Brooke J. System usability scale (SUS): a quick and dirty usability scale. In: Jordan PW, Thomas B, Weerdmeester BA, McClelland IL, editors. Usability evaluation in industry. London (UK): Taylor & Francis; 1996. p. 189–194.

Cain B. A review of the mental workload literature. Toronto (Canada): Human System Integration Section, Defense Research and Development; 2007. RTO-TR-HRM-121-Par-II. Also available at http://www.dtic.mil/dtic/tr/fulltext/u2/a474193.pdf.

Chen JYC, Barnes MJ. Supervisory control of multiple robots: effects of imperfect automation and individual differences. Human Factors. 2012a;54:157–174.

Chen JYC, Barnes MJ. Supervisory control of multiple robots in dynamic tasking environments. Ergonomics. 2012b;55:1043–1058.

Chen JYC, Barnes MJ, Harper-Sciarini M. Supervisory control of multiple robots: Human-performance issues and user-interface design. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions. 2011;41(4):435–454.

Chen JYC, Durlach PJ, Sloan JA, Bowens LD. Human robot interaction in the context of simulated route reconnaissance missions. Military Psychology. 2008;20:135–149.

Chen JYC, Procci K, Boyce M, Wright J, Garcia A, Barnes M. Situation awareness-based agent transparency. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2014 Apr. Report No.: ARL-TR-6905. Also available at: http//www.arl.army.mil/arlreports/2014/ARL-TR-6905.pdf.

Clark K, Fleck MS, Mitroff SR. Enhanced change detection performance reveals improved strategy use in avid action video game players. Acta Psychologica. 2011;136(1):67–72.

Cook M, Smallman H. Human factors of the confirmation bias in intelligence analysis: decision support from graphical evidence landscapes. Human Factors. 2008;50:745–754.

Conway ARA, Kane MJ, Bunting MF, Hambrick ZD, Willhelm O, Engle R. Working memory span tasks: a methodological review and user's guide. Psychonomic Bulletin and Review. 2005;12(5):769–786.

Craig A. Nonparametric measures of sensory efficiency for sustained monitoring tasks. Human Factors: The Journal of the Human Factors and Ergonomics Society. 1979;21(1):69–77.

Cummings ML. Automation and accountability in decision support system interface design. The Journal of Technology Studies. 2006;32(1):23–31.

Cummings ML, Clare A, Hart C. The role of human-automation consensus in multiple unmanned vehicle scheduling. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2010;52(2):348–349.

Davies DR, Parasuraman R. The psychology of vigilance. London (UK): Academic Press; 1982. p. 107–117.

de Visser E, Shaw T, Mohamed-Ameen A, Parasuraman R. Modeling human-automation team performance in networked systems: individual differences in working memory count. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2010;54(14):1087–1091.

Derryberry D, Reed MA. Anxiety-related attentional biases and their regulation by attentional control. Journal of Abnormal Psychology. 2002;111(2):225.

Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP. The role of trust in automation reliance. International Journal of Human-Computer Studies. 2003;58:697–718.

Ekström RB, French JW, Harman HH, Dermen D. Manual for kit of factor referenced cognitive tests. Princeton (NJ): Educational Testing Service; 1976. p. 109–113.

Endsley MR. Toward a theory of situation awareness in dynamic systems. Human Factors. 1995;37(1):32–64.

Fincannon T, Keebler JR, Jentsch F, Curtis M. The influence of camouflage, obstruction, familiarity and spatial ability on target identification from an unmanned ground vehicle. Ergonomics. 2013;56(5):739–751.

Finger R, Bisantz AM. Utilizing graphical formats to convey uncertainty in a decision-making task. Theoretical Issues in Ergonomics Science. 2002;3(1):1–25.

Freedy E, de Visser E, Weltman G, Coeyman N. Measurement of trust in human-robot collaboration. CTS 2007. In: International Symposium on Collaborative Technologies and Systems 2007; 2007 May 25; Orlando, FL. Piscataway (NJ): IEEE; c2007.

Goodrich MA. On maximizing fan-out: towards controlling multiple unmanned vehicles. In: Barnes MG, Jentsch F, editors. Human–robot interactions in future military operations. Surrey (UK): Ashgate; 2010. p. 375–395.

Groom V, Nass C. Can robots be teammates? Benchmarks in human-robot teams. Interaction Studies. 2007;8(3):483–500.

Gugerty L Brooks J. Reference-frame misalignment and cardinal direction judgments: group differences and strategies. Journal of Experimental Psychology: Applied. 2004;10(2):75–88.

Hancock PA, Billings DR, Schaefer KE, Chen JY, De Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Human Factors. 2011;53(5):517–527.

Hart S, Staveland L. Development of NASA TLX (task load index): results of empirical and theoretical research. In: Hancock P, Meshkati N, editors. Human mental workload. Amsterdam (The Netherlands): Elsevier; 1988. p. 139–183.

Hegarty M, Waller D. A dissociation between mental rotation and perspective-taking spatial abilities. Intelligence. 2004;32(2):175–191.

Helldin T, Ohlander U, Falkman G, Riveiro M. Transparency of automated combat classification. In: Harris D, editor. Engineering psychology and cognitive ergonomics. Berlin (Germany): Springer; 2014. p. 22–33.

Hoff KA, Bashir M. Trust in automation integrating empirical evidence on factors that influence trust. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2015;57(3):407–434.

Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H, Van de Weijer J. Eye tracking: a comprehensive guide to methods and measures. New York (NY): Oxford University Press; 2011.

Hwang SL, Yau YJ, Lin YT, Chen JH, Huang TH, Yenn TC, Hsu CC. Predicting work performance in nuclear power plants. Safety Science. 2008;46(7):1115–1124.

Inagaki T, Itoh M. Humans' over-trust in and over-reliance on advanced driver assistance systems: a theoretical framework. International Journal of Vehicular Technology. 2013:1–8.

ISO 9241-171: Ergonomics of human-system interaction – part 171: guidance on software accessibility. Geneva (Switzerland): International Organization for Standardization; 2008.

Jian J, Bisantz AM, Drury CG. Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics. 2000;4(1):53–71.

Kato Y, Takeuchi Y. Individual differences in wayfinding strategies. Journal of Environmental Psychology. 2003;23:171–188.

Lathan CE, Tracey M. The effects of operator spatial perception and sensory feedback on human-robot teleoperation performance. Presence: Teleoperators and Virtual Environments. 2002;11(4):368-377.

Lee JD, See KA. Trust in technology: designing for appropriate reliance. Human Factors. 2004;46:50–80.

Levine TR, Hullett CR. Eta squared, partial eta squared, and misreporting of effect size in communication research. Human Communication Research. 2002;28(4):612–625.

Linegang M, Stoner HA, Patterson MJ, Seppelt BD, Hoffman JD, Crittendon ZB, Lee JD. Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. Proceedings of the 50th Human Factors and Ergonomics Society (HFES) Annual Meeting; 2006 Oct 16–20; San Francisco, CA. Santa Monica (CA): HFES; c2006. p. 2482–2486.

Long L. Visual spatial abilities in uninhabited ground vehicle task performance during teleoperation and direct line of sight. Presence. 2011;20(5):466–479.

Lyons JB, Havig PR. Transparency in a human-machine context: approaches for fostering shared awareness/intent. In: Shumaker R, Lackey S, editors. Virtual, augmented and mixed reality: designing and developing virtual and augmented environments. Berlin (Germany): Springer; 2014. p. 181–190.

Macmillan NA, Creelman CD. Detection theory: A user's guide. Hillsdale (NJ): Lawrence Erlbaum Associates, Inc.; 2004.

McBride M, Morgan S. Trust calibration for automated decision aids. Durham (NC): Institute for Homeland Security Solutions; 2010. [accessed 2015 June 11].

https://www.ihssnc.org/portals/0/Documents/VIMSDocuments/McBride_Research_Brief.pdf.

Miller C, Goldman R, Funk H, Wu P, Pate B. A playbook approach to variable autonomy control: application for control of multiple, heterogeneous unmanned air vehicles. Presented at the Annual Meeting of the American Helicopter Society; 2004 June 7–10; Baltimore, MD.

Neyedli, H, Hollands J, Jamieson G. Beyond identity: incorporating system reliability information into an automated combat identification system. Human Factors. 2011;53:338–355.

Oduor KF, Wiebe EN. The effects of automated decision algorithm modality and transparency on reported trust and task performance. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2008;52(4):302–306.

Paas FG, Van Merriënboer JJ. Instructional control of cognitive load in the training of complex cognitive tasks. Educational Psychology Review. 1994;6(4):351–371.

Parasuraman R, Molloy R, Singh IL. Performance consequences of automation induced “complacency”. The International Journal of Aviation Psychology. 1993;3(1):1–23.

Parasuraman R, Riley V. Humans and automation: use, misuse, disuse, abuse. Human Factors: The Journal of the Human Factors and Ergonomics Society. 1997;39(2):230–253.

Parasuraman R, Sheridan TB, Wickens CD. A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans. 2000;30:286–297.

Pollack I, Norman DA. A non-parametric analysis of recognition experiments. Psychonomic Science. 1964;1(1–12):125–126.

Quayle JD, Ball LJ. Working memory, metacognitive uncertainty, and belief bias in syllogistic reasoning. The Quarterly Journal of Experimental Psychology: Section A. 2000;53(4):1202–1223.

Rao AS, Georgeff MP. BDI-agents: from theory to practice. Proceedings of the First International Conference on Multiagent Systems; 1995 June 12–14; San Francisco, CA. San Francisco (CA): Association for the Advancement of Artificial Intelligence (AAAI) Press; c1995. p. 312–319.

Russell SJ, Norvig P. Artificial intelligence: a modern approach, 3rd edition. Saddle River (NJ): Prentice Hall; 2009.

Sarter NB, Woods DD. How in the world did we ever get into that mode? Mode error and awareness in supervisory control. Human Factors: The Journal of the Human Factors and Ergonomics Society. 1995;37(1):5–19.

Scholtz J, Consolvo S. Toward a framework for evaluating ubiquitous computing applications. Pervasive Computing. 2004;3(2):82–88.

Spriggs, S, Boyer, J, Bearden, G. Fusion system overview. Lecture conducted at Wright Patterson Air Force Base, Ohio; 2014.

Van Orden KF, Limbert W, Makeig S, Jung TP. Eye activity correlates of workload during a visuospatial memory task. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2001:43(1):111–121.

Wang L, Jamieson GA, Hollands JG. Trust and reliance on an automated combat identification system. Human Factors. 2009;51:281–291.

Wickens CD. Multiple resources and mental workload. Human Factors: The Journal of the Human Factors and Ergonomics Society. 2008;50(3):449–455.

Wickens CD, Dixon SR. The benefits of imperfect diagnostic automation: a synthesis of the literature. Theoretical Issues in Ergonomics Science. 2007;8(3):201–212.

Williams EJ. Experimental designs balanced for the estimation of residual effects of treatments. Australian Journal of Chemistry. 1949;2(2):149–168.

Wright JL, Chen JY, Quinn SA, Barnes MJ. The effects of level of autonomy on human-agent teaming for multi-robot control and local security maintenance. Aberdeen Proving Ground (MD): Army Research Laboratory (US); 2013 Nov. Report No.: ARL-TR-6724. Also available at: http://www.arl.army.mil/arlreports/2013/ARL-TR-6724.pdf.

Yang C, Liu Z, Zhou Q, Xie F, and Zhou S. Analysis on eye movement indexes based on simulated flight task. Proceedings of Engineering Psychology and Cognitive Ergonomics 2014; 2014 June 22–27; Heraklion, Crete. p. 419-427.

Zaichkowsky JL. Measuring the involvement construct. Journal of Consumer Research. 1985;12(3):341–342.

INTENTIONALLY LEFT BLANK.

Appendix A. Demographics

This appendix appears in its original form, without editorial change.

Demographic Questionnaire

**Participant # _______ Age ______ Major ________________ Date
___________ Gender ___**

1. What is the <u>highest</u> level of education you have had? (*<u>Circle one only</u>*)
a) Less than 4 yrs of college b) Completed 4 yrs of college c) Other

2. When did you use computers in your education? (*<u>Check all that apply</u>*)

❐ Grade School	❐ High School	❐ College
❐ Jr. High	❐ Technical School	❐ Did Not Use

3. Where do you currently use a computer? (*<u>Check all that apply</u>*)

❐ Home
❐ Work
❐ Library
❐ Other_______
❐ Do Not Use

4. For each of the following questions, circle the response that best describes you.

How often do you:

Use a mouse?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Use a joystick?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Use a touch screen?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Use icon-based programs/software?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Use programs/software with pull-down menus?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Use graphics/drawing features in software packages?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Use E-Mail?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Operate a radio controlled vehicle (car, boat, or plane)?	Daily	Weekly	Monthly	Once every few months	Rarely	Never
Play computer/video games?	Daily	Weekly	Monthly	Once every few months	Rarely	Never

5. Which type(s) of computer/video games do you most often play if you play at least once every few months?:

6. Which of the following best describes your expertise with computers?
(*Circle one only*)

a) Novice
b) Good with one type of software package (such as word processing or slides)
c) Good with several software packages
d) Can program in one language and use several software packages
e) Can program in several languages and use several software packages

7. Are you in your good/ comfortable state of health physically? YES NO

If NO, please briefly explain:

8. How many hours of sleep did you get last night? ______ hours

9. Do you have normal color vision? YES NO

10. Do you have prior military service? YES NO

If YES, how long?: __________ years

Please answer the following questions about how you play video games by circling a number on the provided scale, from **1** (**strongly disagree**) to **6** (**strongly agree**).

	Strongly Disagree					**Strongly Agree**
11. I can always manage to solve difficult problems within a video game if I try hard enough.	**1**	**2**	**3**	**4**	**5**	**6**
12. In a video game, if someone opposes me, I can find the means and ways to get what I want.	**1**	**2**	**3**	**4**	**5**	**6**
13. It is easy for me to stick to my plans and accomplish my goals in a video game.	**1**	**2**	**3**	**4**	**5**	**6**
14. I am confident that I could deal efficiently with unexpected events in a video game.	**1**	**2**	**3**	**4**	**5**	**6**
15. Thanks to my resourcefulness, I know how to handle unforeseen situations in a video game.	**1**	**2**	**3**	**4**	**5**	**6**

16. I can solve most problems in a video game if I invest the necessary effort.	**1**	**2**	**3**	**4**	**5**	**6**
17. I can remain calm when facing difficulties in a video game because I can rely on my coping abilities.	**1**	**2**	**3**	**4**	**5**	**6**
18. When I am confronted with a problem in a video game, I can usually find several solutions.	**1**	**2**	**3**	**4**	**5**	**6**
19. If I am in trouble in a video game, I can usually think of a solution.	**1**	**2**	**3**	**4**	**5**	**6**
20. I can usually handle whatever comes my way in a video game.	**1**	**2**	**3**	**4**	**5**	**6**

Please answer the following questions about how you feel about automation by circling number on the provided scale, from **1** (**strongly disagree**) to **5** (**strongly agree**).

	strongly disagree				**strongly agree**
1. I usually trust automation until there is a reason not to.	**1**	**2**	**3**	**4**	**5**
2. For the most part, I DISTRUST automation.	**1**	**2**	**3**	**4**	**5**
3. In general, I would rely on automation to assist me.	**1**	**2**	**3**	**4**	**5**
4. My tendency to trust automation is high.	**1**	**2**	**3**	**4**	**5**
5. It is easy for me to trust automation to do its job.	**1**	**2**	**3**	**4**	**5**
6. I am likely to trust automation even when I have little knowledge about it.	**1**	**2**	**3**	**4**	**5**

INTENTIONALLY LEFT BLANK.

Appendix B. Trust Scale

This appendix appears in its original form, without editorial change.

Trust Survey

For each of the following items and situations, circle the number which best describes your feeling or your impression based on the system you just used. For each item, consider the following situations:

- A: When the system is collecting and/or highlighting/filtering information.
- B: When the system is integrating information, generating predictive displays, and/or presenting its analysis.
- C: When the system is making decisions and/or selecting actions.
- D: When the system is executing actions.

1. The system is deceptive when…

			not at all				
	neutral			*extremely*			
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

2. The system behaves in an underhanded manner when…

			not at all				
		neutral				*extremely*	
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

3. I am suspicious of the system's intent, action, or outputs when…

			not at all				
		neutral				*extremely*	
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

4. I am wary of the system when…

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

5. The system's actions will have a harmful or injurious outcome when…

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

6. I am confident in the system when…

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

7. The system provides security when…

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

8. The system has integrity when…

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7

B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

9. The system is dependable when…

	not at all						
	neutral				*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

10. The system is reliable when…

	not at all						
	neutral				*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

11. I can trust the system when…

	not at all						
	neutral				*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

12. I am familiar with the system when…

	not at all						
	neutral				*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

13. The system is predictable when...

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

14. The system meets the needs of the mission when...

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

15. The system provides appropriate information when...

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

16. The system malfunctions when...

		not at all					
		neutral			*extremely*		
A: Gathering or Filtering Information	1	2	3	4	5	6	7
B: Integrating and Displaying Analyzed Information	1	2	3	4	5	6	7
C: Suggesting or Making Decisions	1	2	3	4	5	6	7
D: Executing Actions	1	2	3	4	5	6	7

Now imagine that you are employed as an unmanned vehicle operator to complete missions. Reflecting on the experience with the system you just used, please rate the extent to which you agree with each of these items by circling a value from **1** (**strongly disagree**) to **7** (**strongly agree**), where **4** is **neutral**.

	Strongly Disagree			**Neutral**			**Strongly Agree**
17. Using the system would improve my job performance.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
18. Using the system would make it easier to do my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
19. I would find the system useful in my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
20. Learning to operate the system is easy for me.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
21. It is easy for me to become skillful at using the system.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
22. I find the system easy to use.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
23. I intend to use this system for my job.	**1**	**2**	**3**	**4**	**5**	**6**	**7**

Appendix C. National Air and Space Administration Task Load Index (NASA-TLX)

This appendix appears in its original form, without editorial change.

NASA-TLX Questionnaire

Please rate your overall impression of demands imposed on you during the exercise.

1. Mental Demand: How much mental and perceptual activity was required (e.g., thinking, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?

LOW |---|---|---|---|---|---|---|---|---| HIGH
1 2 3 4 5 6 7 8 9 10

2. Physical Demand: How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?

LOW |---|---|---|---|---|---|---|---|---| HIGH
1 2 3 4 5 6 7 8 9 10

3. Temporal Demand: How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?

LOW |---|---|---|---|---|---|---|---|---| HIGH
1 2 3 4 5 6 7 8 9 10

4. Level of Effort: How hard did you have to work (mentally and physically) to accomplish your level of performance?

LOW |---|---|---|---|---|---|---|---|---| HIGH
1 2 3 4 5 6 7 8 9 10

5. Level of Frustration: How insecure, discouraged, irritated, stressed and annoyed versus secure, gratified, content, relaxed and complacent did you feel during the task?

LOW |---|---|---|---|---|---|---|---|---| HIGH
1 2 3 4 5 6 7 8 9 10

6. Performance: How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?

LOW |---|---|---|---|---|---|---|---|---| HIGH
1 2 3 4 5 6 7 8 9 10

Pairwise Comparison of Factors

Select the member of each pair that provided the most significant source of workload variation in these tasks.

Physical Demand vs. Mental Demand

Temporal Demand vs. Mental Demand

Performance vs. Mental Demand

Frustration vs. Mental Demand

Effort vs. Mental Demand

Temporal Demand vs. Physical Demand

Performance vs. Physical Demand

Frustration vs. Physical Demand

Effort vs. Physical Demand

Temporal Demand vs. Performance

Temporal Demand vs. Frustration

Temporal Demand vs. Effort

Performance vs. Frustration

Performance vs. Effort

Effort vs. Frustration

INTENTIONALLY LEFT BLANK.

Appendix D. System Usability Scale

This appendix appears in its original form, without editorial change.

System Usability Scale

Please answer the following questions about the system you just used by circling a number on the provided response scale, from **1** (**strongly disagree**) to **5** (**strongly agree**).

	Strongly Disagree				**Strongly Agree**
1. I think that I would like to use this system frequently.	**1**	**2**	**3**	**4**	**5**
2. I found the system unnecessarily complex.	**1**	**2**	**3**	**4**	**5**
3. I thought the system was easy to use.	**1**	**2**	**3**	**4**	**5**
4. I think that I would need the support of a technical person to be able to use this system.	**1**	**2**	**3**	**4**	**5**
5. I found the various functions in this system were well integrated.	**1**	**2**	**3**	**4**	**5**
6. I thought there was too much inconsistency in this system.	**1**	**2**	**3**	**4**	**5**
7. I would imagine that most people would learn to use this system very quickly.	**1**	**2**	**3**	**4**	**5**
8. I found the system very awkward to use.	**1**	**2**	**3**	**4**	**5**
9. I felt very confident using the system.	**1**	**2**	**3**	**4**	**5**
10. I needed to learn a lot of things before I could get going with this system.	**1**	**2**	**3**	**4**	**5**

Appendix E. Attentional Control Scale

This appendix appears in its original form, without editorial change.

Attentional Control Survey

For each of the following questions, circle the response that best describes you.

It is very hard for me to concentrate on a difficult task when there are noises around.	**Almost Never**	**Sometimes**	**Often**	**Always**
When I need to concentrate and solve a problem, I have trouble focusing my attention.	**Almost Never**	**Sometimes**	**Often**	**Always**
When I am working hard on something, I still get distracted by events around me.	**Almost Never**	**Sometimes**	**Often**	**Always**
My concentration is good even if there is music in the room around me.	**Almost Never**	**Sometimes**	**Often**	**Always**
When concentrating, I can focus my attention so that I become unaware of what's going on in the room around me.	**Almost Never**	**Sometimes**	**Often**	**Always**
When I am reading or studying, I am easily distracted if there are people talking in the same room.	**Almost Never**	**Sometimes**	**Often**	**Always**
When trying to focus my attention on something, I have difficulty blocking out distracting thoughts.	**Almost Never**	**Sometimes**	**Often**	**Always**
I have a hard time concentrating when I'm excited about something.	**Almost Never**	**Sometimes**	**Often**	**Always**
When concentrating, I ignore feelings of hunger or thirst.	**Almost Never**	**Sometimes**	**Often**	**Always**
I can quickly switch from one task to another.	**Almost Never**	**Sometimes**	**Often**	**Always**
It takes me a while to get really involved in a new task.	**Almost Never**	**Sometimes**	**Often**	**Always**
It is difficult for me to coordinate my attention between the listening and writing required when taking notes during lectures.	**Almost Never**	**Sometimes**	**Often**	**Always**
I can become interested in a new topic very quickly when I need to.	**Almost Never**	**Sometimes**	**Often**	**Always**
It is easy for me to read or write while I'm also talking on the phone.	**Almost Never**	**Sometimes**	**Often**	**Always**
I have trouble carrying on two conversations at once.	**Almost Never**	**Sometimes**	**Often**	**Always**
I have a hard time coming up with new ideas quickly.	**Almost Never**	**Sometimes**	**Often**	**Always**
After being interrupted or distracted, I can easily shift my attention back to what I was doing before.	**Almost Never**	**Sometimes**	**Often**	**Always**
When a distracting thought comes to mind, it is easy for me to shift my attention away from it.	**Almost Never**	**Sometimes**	**Often**	**Always**

It is easy for me to alternate between two different tasks.	**Almost Never**	**Sometimes**	**Often**	**Always**
It is hard for me to break from one way of thinking about something and look at it from another point of view.	**Almost Never**	**Sometimes**	**Often**	**Always**

INTENTIONALLY LEFT BLANK.

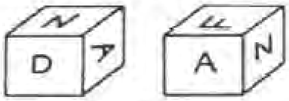
Appendix F. Cube Comparison Test

This appendix appears in its original form, without editorial change.

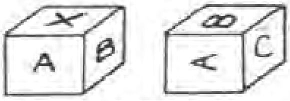
Cube Comparisons Test **Participant # ________ Date ______**

CUBE COMPARISONS TEST -- S-2 (Rev.)

Wooden blocks such as children play with are often cubical with a different letter, number, or symbol on each of the six faces (top, bottom, four sides). Each problem in this test consists of drawings of pairs of cubes or blocks of this kind. Remember, there is a different design, number, or letter on each face of a given cube or block. Compare the two cubes in each pair below.



S □ D ■



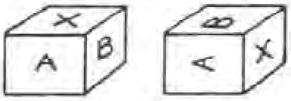
S ■ D □

The first pair is marked D because they must be drawings of different cubes. If the left cube is turned so that the A is upright and facing you, the N would be to the left of the A and hidden, not to the right of the A as is shown on the right hand member of the pair. Thus, the drawings must be of different cubes.

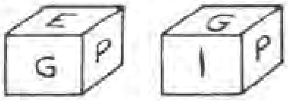
The second pair is marked S because they could be drawings of the same cube. That is, if the A is turned on its side the X becomes hidden, the B is now on top, and the C (which was hidden) now appears. Thus the two drawings could be of the same cube.

Note: No letters, numbers, or symbols appear on more than one face of a given cube. Except for that, any letter, number or symbol can be on the hidden faces of a cube.

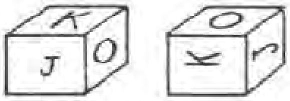
Work the three examples below.



S □ D □



S □ D □



S □ D □

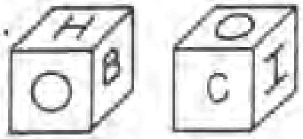
The first pair immediately above should be marked D because the X cannot be at the peak of the A on the left hand drawing and at the base of the A on the right hand drawing. The second pair is "different" because P has its side next to G on the left hand cube but its top next to G on the right hand cube. The blocks in the third pair are the same, the J and K are just turned on their side, moving the O to the top.

Your score on this test will be the number marked correctly minus the number marked incorrectly. Therefore, it will not be to your advantage to guess unless you have some idea which choice is correct. Work as quickly as you can without sacrificing accuracy.

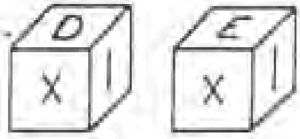
You will have 3 minutes for each of the two parts of this test. Each part has one page. When you have finished Part 1, STOP.

DO NOT TURN THE PAGE UNTIL YOU ARE ASKED TO DO SO.

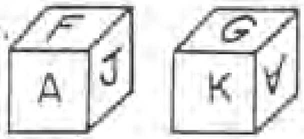
Part 1 (3 minutes)

1.


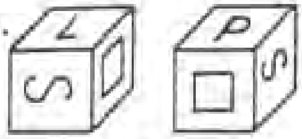
S □ D □

2.


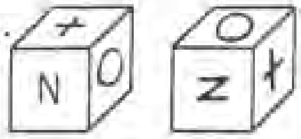
S □ D □

3.


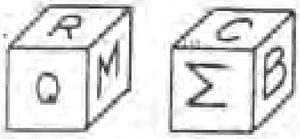
S □ D □

4.


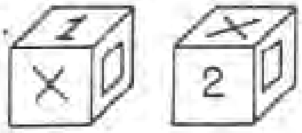
S □ D □

5.


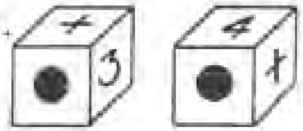
S □ D □

6.


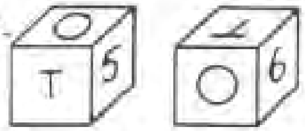
S □ D □

7.


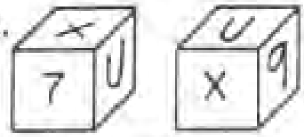
S □ D □

8.


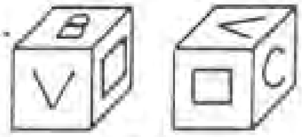
S □ D □

9.


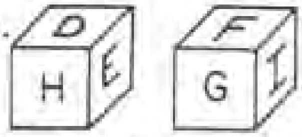
S □ D □

10.


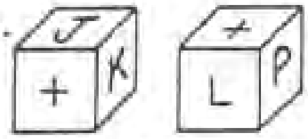
S □ D □

11.


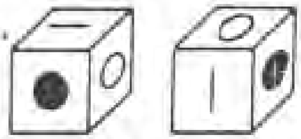
S □ D □

12.


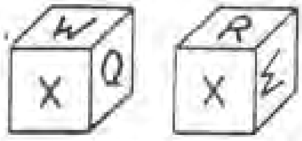
S □ D □

13.


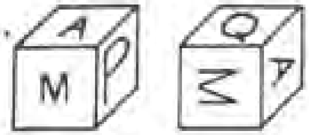
S □ D □

14.


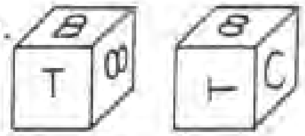
S □ D □

15.


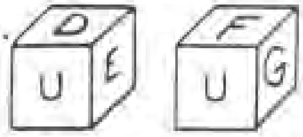
S □ D □

16.


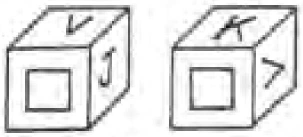
S □ D □

17.


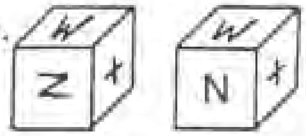
S □ D □

18.


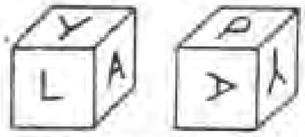
S □ D □

19.


S □ D □

20.


S □ D □

21.


S □ D □

DO NOT GO ON TO THE NEXT PAGE UNTIL ASKED TO DO SO. STOP.

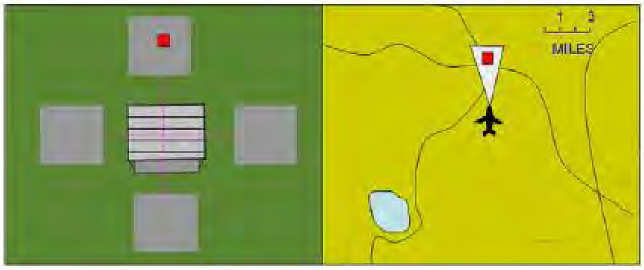
INTENTIONALLY LEFT BLANK.

Appendix G. Spatial Orientation Test

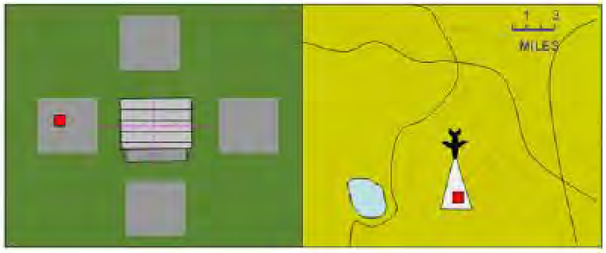
This appendix appears in its original form, without editorial change.

Spatial Orientation Test

The Spatial Orientation Test, modeled after the cardinal direction test developed by Gugerty and his colleagues (Gugerty & Brooks, 2004), is a computerized test consisting of a brief training segment and 32 test questions. The program automatically captures both accuracy and response time. Participants are shown the following image:



The right side image is of a map showing a plane flying. The left side of the display is the pilot's view (from the cockpit of the plane) of several parking lots surrounding a building. The participants' task is to use the right side of the display to learn in which direction the plane is flying. They then use this information to identify which parking lot (north, south, east, or west) in the left side image has the dot. In the example shown above, the plane is heading north, and so the dot appears in the north parking lot. In the example shown below, the plane is heading south, and so the dot appears in the east parking lot.



Participants are shown 32 of these images in succession; each time the direction the plane is flying and the location of the dot are randomized. Participants answer by clicking on one of four buttons (North, South, East, or West). This test is self-paced; the participant may take as long as they wish to answer, and when they answer one question the next question automatically appears. No questions can be skipped, and the order of images is randomized among participants.

Appendix H. Sense of Direction Questionnaire

This appendix appears in its original form, without editorial change.

SDQ-S

Please answer the following questions by circling a number on the provided response scale, from **1** (**strongly disagree**) to **5** (**strongly agree**).

	Strongly Disagree	**Disagree**	**Neutral**	**Agree**	**Strongly Agree**
1. I can make correct choices as to cardinal directions in an unfamiliar place.	**1**	**2**	**3**	**4**	**5**
2. I have become confused, as to cardinal directions, when I am in an unfamiliar place.	**1**	**2**	**3**	**4**	**5**
3. I have difficulties identifying the moving direction of a train with regard to cardinal direction.	**1**	**2**	**3**	**4**	**5**
4. When I get route information, I can make use of ‘‘left or right’’ information, but I can’t use cardinal directions.	**1**	**2**	**3**	**4**	**5**
5. I can’t make out which direction my room in a hotel faces.	**1**	**2**	**3**	**4**	**5**
6. I can tell where I am on a map.	**1**	**2**	**3**	**4**	**5**
7. I can visualize the route as a map-like image.	**1**	**2**	**3**	**4**	**5**
8. I feel anxious about my walking direction in an unfamiliar area.	**1**	**2**	**3**	**4**	**5**
9. I have poor memory for landmarks.	**1**	**2**	**3**	**4**	**5**
10. I cannot remember landmarks found in the area where I have often been.	**1**	**2**	**3**	**4**	**5**
11. I can’t use landmarks in wayfinding.	**1**	**2**	**3**	**4**	**5**
12. I can’t remember the different aspects of sceneries.	**1**	**2**	**3**	**4**	**5**

13. I often can’t find the way even if given detailed verbal information on the route.	**1**	**2**	**3**	**4**	**5**
14. I have a lot of difficulties reaching the unknown place even after looking at a map.	**1**	**2**	**3**	**4**	**5**
15. I often (or easily) forget which direction I turned.	**1**	**2**	**3**	**4**	**5**
16. I become totally confused as to the correct sequence of the return way as a consequence of a number of left-right turns in the route.	**1**	**2**	**3**	**4**	**5**
17. I can’t verify landmarks in a turn of the route.	**1**	**2**	**3**	**4**	**5**

INTENTIONALLY LEFT BLANK.

Appendix I. Personal Involvement Measure

This appendix appears in its original form, without editorial change.

Personal Involvement Measure

Reflecting on the experience with the system you just used, please rate the extent to which you agree with each of these items by circling a value from **1** (**strongly disagree**) to **7** (**strongly agree**), where **4** is **neutral**.

	Strongly Disagree			**Neutral**			**Strongly Agree**
1. I was uninterested in the task.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
2. Doing well in the task was important to me.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
3. The task was trivial.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
4. The task mattered to me.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
5. I was motivated to do the task.	**1**	**2**	**3**	**4**	**5**	**6**	**7**
6. I was unconcerned with doing well in the task.	**1**	**2**	**3**	**4**	**5**	**6**	**7**

Appendix J. Structured Strategy Interview Questions

This appendix appears in its original form, without editorial change.

Structured Strategy Interview Questions

Part 1.

Please answer the following questions regarding the last set of decisions you have made. When you answer only think of the missions you just completed.

1. Can you describe the overall process you used to make a decision? How did you decide which plan to choose between the two plans that you were presented with?

2. Overall, which parts of the system or display elements did you consider when making a decision? List all the parts that you used.

3. Were there any decisions that you relied more on certain parts of the system than others? Or any that certain parts of the system were not helpful at all to your decision making process?

4. If you only had one piece of information from the system to use to solve all of the decisions you encountered which one would you want to use?

Part 2.

Previously you mentioned several parts of the system that you used to make a decision. Now you are going to rate each part of the system on a 1-7 scale. A rating of a 1 indicates this part was not helpful at all, while a rating of a 7 indicates it was extremely helpful to your decision making process.

1. Play name (what the play was called)	1	2	3	4	5	6	7
2. Play details tile	1	2	3	4	5	6	7
3. Vehicle status indicator	1	2	3	4	5	6	7
4. Information Bar	1	2	3	4	5	6	7
5. Plan colors (colors of vehicles and map elements)	1	2	3	4	5	6	7
6. Asset capability display	1	2	3	4	5	6	7
7. Vehicle Sizes	1	2	3	4	5	6	7
8. Map (locations of icons on map, vehicles etc…)	1	2	3	4	5	6	7
9. Intel Alerts	1	2	3	4	5	6	7
10. Equalizer Display	1	2	3	4	5	6	7
11. Text Table	1	2	3	4	5	6	7
12. Equalizer display uncertainty	1	2	3	4	5	6	7
13. Table Uncertainty	1	2	3	4	5	6	7
14. Vehicle Uncertainty	1	2	3	4	5	6	7
15. Vehicle path uncertainty	1	2	3	4	5	6	7

Part 3.

1. In today's experiment you had three different system layouts with different types of Information. Which one did you prefer? Were decisions easier with one layout than the others?

2. Were there any parts of the system that gave you consistency, conflicting or hard to understand Information? Please be as detailed as possible.

3. Do you have any other comments for us about the experiment?

INTENTIONALLY LEFT BLANK.

List of Symbols, Abbreviations, and Acronyms

3Ps	purpose, process, performance
AFRL	US Air Force Research Laboratory
ANOVA	analysis of variance
AVGP	action video game player
CI	confidence interval
CR	correct IA rejection
FD	fixation duration
GE	gaming experience
IA	intelligent agent
ID	individual difference
Intel	intelligence
ISO	International Organization for Standardization
MANOVA	multivariate analysis of variance
NASA-TLX	National Air and Space Administration Task Load Index
OSPAN	Operation Span
PAC	perceived attentional control
PD	pupil diameter
PU	proper IA use
RT	response time
SA	situation awareness
SAT	SA-based agent transparency
SDT	signal detection theory
SEM	standard error of the mean
SMI RED	SensoMotoric Instruments Remote Eye-tracking Device
SpaO	spatial orientation

SpaV	spatial visualization
SUS	System Usability Scale
UAV	unmanned aerial vehicle
UGV	unmanned ground vehicle
USV	unmanned surface vehicle
UxV	multi-unmanned vehicle
WMC	working memory capacity

1 DEFENSE TECHNICAL
(PDF) INFORMATION CTR
DTIC OCA

2 DIRECTOR
(PDF) US ARMY RESEARCH LAB
RDRL CIO LL
IMAL HRA MAIL AND
RECORDS MGMT

1 GOVT PRINTG OFC
(PDF) A MALHOTRA

1 DIR US ARMY EVALCTR HQ
(HC) TEAE SV
P A THOMPSON
2202 ABERDEEN BLVD 2ND FL
APG MD 21005-5001

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM D
T DAVIS
BLDG 5400 RM C242
REDSTONE ARSENAL AL
35898-7290

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRS EA DR V J RICE
BLDG 4011 RM 217
1750 GREELEY RD
FORT SAM HOUSTON TX
78234-5002

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM DG K GUNN
BLDG 333
PICATINNY ARSENAL NJ
07806-5000

1 ARMY RSCH LAB – HRED
(PDF) ARMC FIELD ELEMENT
RDRL HRM CH C BURNS
THIRD AVE
BLDG 1467B RM 336
FORT KNOX KY 40121

1 ARMY RSCH LAB – HRED
(PDF) AWC FIELD ELEMENT
RDRL HRM DJ D DURBIN
BLDG 4506 (DCD) RM 107
FORT RUCKER AL 36362-5000

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM CK J REINHART
10125 KINGMAN RD BLDG 317
FORT BELVOIR VA 22060-5828

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM AY M BARNES
2520 HEALY AVE
STE 1172 BLDG 51005
FORT HUACHUCA AZ
85613-7069

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM AP D UNGVARSKY
POPE HALL BLDG 470
BCBL 806 HARRISON DR
FORT LEAVENWORTH KS
66027-2302

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM AR J CHEN
12423 RESEARCH PKWY
ORLANDO FL 32826-3276

1 ARMY RSCH LAB – HRED
(PDF) HUMAN SYSTEMS
INTEGRATION ENGR
TACOM FIELD ELEMENT
RDRL HRM CU P MUNYA
6501 E 11 MILE RD
MS 284 BLDG 200A
WARREN MI 48397-5000

1 ARMY RSCH LAB – HRED
(PDF) FIRES CTR OF EXCELLENCE
FIELD ELEMENT
RDRL HRM AF C HERNANDEZ
3040 NW AUSTIN RD RM 221
FORT SILL OK 73503-9043

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM AV W CULBERTSON
91012 STATION AVE
FORT HOOD TX 76544-5073

8 ARMY RSCH LAB – HRED
(PDF) SIMULATION & TRAINING
TECHNOLOGY CENTER
RDRL HRT COL G LAASE
RDRL HRT I MARTINEZ
RDRL HRT T R SOTTILARE
RDRL HRT B N FINKELSTEIN
RDRL HRT G A RODRIGUEZ
RDRL HRT I J HART
RDRL HRT M C METEVIER
RDRL HRT S B PETTIT
12423 RESEARCH PARKWAY
ORLANDO FL 32826

1 ARMY RSCH LAB – HRED
(PDF) RDRL HRM DE A MARES
1733 PLEASONTON RD BOX 3
FORT BLISS TX 79916-6816

1 ARMY RSCH LAB – HRED
(PDF) HQ USASOC
RDRL HRM CN R SPENCER
BLDG E2929 DESERT STORM DR
FORT BRAGG NC 28310

1 ARMY G1
(PDF) DAPE MR B KNAPP
300 ARMY PENTAGON
RM 2C489
WASHINGTON DC 20310-0300

<u>ABERDEEN PROVING GROUND</u>

17 DIR USARL
(PDF) RDRL HR
L ALLENDER
P FRANASZCZUK
K MCDOWELL
RDRL HRM
P SAVAGE-KNEPSHIELD
RDRL HRM AL
C PAULILLO
RDRL HRM AR
J MERCADO
RDRL HRM AT
J CHEN
M RUPP
RDRL HRM AY
M BARNES
RDRL HRM B
J GRYNOVICKI
RDRL HRM C
L GARRETT
RDRL HRS
J LOCKETT
RDRL HRS B
M LAFIANDRA
RDRL HRS D
A SCHARINE
RDRL HRS E
D HEADLEY
RDRL SL
D BAYLOR
RDRL SLE
R FLORES